House Bill 1164 Writing Pilot Program

Report to the Governor and The Texas Legislature

9/1/2017

Texas Education Agency

Table of Contents

WRITING PILOT BACKGROUND

As required by House Bill (HB) 1164, 84th Texas Legislature, 2015, for the 2016–2017 and 2017–2018 school years, TEA and Educational Testing Service (ETS) will conduct a pilot program study to examine alternative methods of assessing writing.

The pilot study will include the collection and scoring of a range of student writing samples produced throughout the school year. The writing products to be completed, submitted, and scored are:

- two timed writing samples completed at the beginning and end of the school year based on a specific writing prompt chosen by each student from a selection of three prompts;
- three instructional writing process samples from different genres (i.e., personal narrative, expository, persuasive, or analytic) that include evidence of a writing process from start to finish (e.g., planning, drafting, revising, editing, and publishing); and
- an instructional portfolio containing the writing samples listed above.

Scoring of the student writing samples consists of several components. The student samples will initially be scored by each student's teacher of record. Additionally, the samples will receive a second blind score that will be coordinated at the local level by each participating Education Service Center (ESC) and include local teachers who are certified to teach English language arts. A final sampling of scores will be conducted by TEA and ETS.

YEAR-ONE OVERVIEW

For the pilot study, three regional ESCs were selected to participate with a total of seven partnering school districts for year one. Region 6 (Huntsville) partnered with Calvert ISD and Huntsville ISD. Region 10 (Richardson) partnered with Athens ISD, Garland ISD, and Sunnyvale ISD. Region 16 (Amarillo) partnered with Amarillo ISD and Dumas ISD. In total, 37 teachers and 1,707 students in grade 4, grade 7, English I, and English II from across the state of Texas participated in year one of the writing pilot.

The 2016–2017 school year began with English language arts and reading (ELA/R) representatives from the partnering ESCs attending a kick-off planning session with TEA and ETS in Austin. During the daylong collaboration opportunity, the writing specialists set goals, established timelines, decided on timed writing sample prompts, and developed the foundation of the writing pilot rubric (see Appendix A).

Once the writing pilot rubric was established, a companion scoring training was developed to introduce participating teachers to using the rubric to assess student writing. TEA and ETS then facilitated a virtual train-the-trainer session for the three regional ESC representatives who, in turn, held in-person scoring trainings for the participating teachers in their region.

Communication and collaboration remained a high priority during year one. Representatives from TEA, the ESCs, and ETS met weekly to plan and monitor pilot program activities. In addition to the weekly

meetings, both TEA and ETS were available for one-on-one support to any ESC, district, and/or teacher who needed assistance. In this collaborative method, a series of ongoing resources and the portfolio scoring rubric (see Appendix B) and its training were developed.

SAMPLES

To establish a baseline of student writing, Timed Writing Sample 1 was conducted at the end of September 2016. Students were given an in-class timed writing assignment and had the opportunity to choose from three prompts to write about within the given time period. While there was a time restriction (see chart below), there was no length restriction. Students were free to write as much as they wanted to within the given time.

Grade/Course	**Time Limit**
Grade 4	35 minutes
Grade 8	45 minutes
English I and English II	60 minutes

During the fall and spring semesters, teachers participating in the writing pilot worked on the instructionally based writing process samples with their students. The process samples were assigned and collected according to the appropriate grade-level genres outlined in the Texas Essential Knowledge and Skills standards (TEKS), specifically personal narrative, expository, persuasive, and analytic. These samples, along with both timed samples, were compiled into a student's writing portfolio and contained evidence of the student's writing process (e.g., planning, drafting, revising, editing, and publishing).

Teachers were provided with designated timeframes and submission windows for assigning and collecting each writing process sample. Participating districts and teachers could choose the writing genre to collect during each submission window. Submission windows and choice of genre gave teachers the flexibility to fully align the assessment with local instruction and scope and sequence of curriculum. In addition, to better support districts in their writing instruction scope and sequence, a decision was made mid-year by the pilot leadership team to collect two rather than three writing process samples.

Timed Writing Sample 2 was assigned during the last two weeks of April 2017. Students were given a choice of three prompts to write about and the same time allotment and genre as Timed Writing Sample 1.

During year one of the pilot, the student samples were collected and housed according to the decision of each local district. Some teachers asked their students to work on a computer for their assignments while others asked their students to complete the assignments on paper. In addition, some teachers housed their student portfolios in accordion files, binders, or folders, while others stored their student portfolios digitally. All samples to be scored for year one were periodically uploaded throughout the year to a secure online file and stored in a secure database.

YEAR-ONE SCORING

Classroom teachers scored the writing pilot samples at varying times throughout the school year using two rubrics—the writing pilot rubric and the portfolio scoring rubric. With the writing pilot rubric, classroom teachers scored the students' Timed Writing Sample 1 assignments, the final copy of the writing process samples, and Timed Writing Sample 2 assignments upon completion in accordance with the writing pilot scoring deadlines. Towards the end of the school year, classroom teachers scored their students' collected portfolio samples using the portfolio scoring rubric. All teacher-of-record scores, along with student samples, were submitted throughout the year and stored in the secure writing pilot database.

The blind scoring sessions for writing samples were held in June 2017. During the blind scoring sessions, all participating students' writing samples and portfolios were scored at the local ESC level by teachers certified to teach English language arts. Each of the three participating ESCs recruited teachers within their respective regions for the blind scoring. Each regional blind scoring session consisted of three full days. The lead ESC writing pilot representative conducted teacher training on both rubrics using training materials collaboratively developed by representatives from TEA, the ESCs, and ETS. The training sessions lasted approximately three hours for each of the two rubrics. Region 6 (Huntsville) trained 23 teachers; Region 10 (Richardson) trained 31 teachers; and Region 16 (Amarillo) trained 31 teachers. Over the course of the three days, teachers at each regional session scored a random sample of one-third of the statewide writing pilot samples and portfolios. Teachers recorded their identifying numbers and their ratings on score sheets. After the end of the scoring sessions, ETS keyed the scores into spreadsheets to upload into the secure database. All teacher raters completed end-of-scoring-session evaluation surveys providing input on their scoring experience.

A third sampling of scores was conducted by ETS on behalf of TEA during the last week of June 2017. ETS recruited Texas-based experienced raters who were certified for constructed response scoring for the State of Texas Assessments of Academic Assessments (STAAR®). An ETS ELA assessment specialist involved with the writing pilot trained the ETS raters using the same materials used by the ESCs. The training time for each rubric was the same as the training time used with the ESCs, approximately 3 hours. ETS raters scored all complete portfolios using blind scoring—no rater saw any score from other raters for any of the portfolio components. ETS raters used identical score sheets as the teacher raters, and ETS keyed this set of scores into spreadsheets to uploads into the database. All ETS raters completed end-of-scoring-session evaluation surveys providing input on their scoring experience.

YEAR-ONE DATA ANALYSIS

YEAR-ONE DATA COLLECTION OVERVIEW

A stratified random sampling was conducted and classes of students from three ESCs were recruited for the study. A complete writing portfolio included two timed writing samples and two or three process writing samples that were collected from each participating student in the following order across the

school year: Timed Writing Sample 1, Process Writing Sample 1, Process Writing Sample 2, Process Writing Sample 3 (if available), and Timed Writing Sample 2.

Each individual final writing sample (i.e., final copies of timed writing samples and process writing samples) in a student portfolio received a rating score according to the writing pilot rubric. In addition, each complete student portfolio received seven rating scores according to the writing pilot portfolio rubric on (1) Planning, (2) Drafting, (3) Revising, (4) Editing, Publishing, and Attention to Feedback, (5) Expressing Ideas, (6) Organization and Structure, and (7) Use of Language and Conventions.

Each student's writing samples and portfolio were independently rated according to the corresponding rubrics by three types of raters: (1) the classroom teacher of record, (2) a rater selected by the ESC, and (3) a trained rater (i.e., a rater previously certified to score the STAAR constructed responses who was recruited and trained by ETS on the writing pilot rubrics). Additionally, 20%–30% of the students' writing samples and portfolios received ratings from a second group of trained raters for the purpose of studying the quality of ratings assigned by the first group of trained raters. A teacher rated a writing sample right after it was collected, whereas the other raters blindly rated only the complete portfolio after all writing samples in the portfolio were collected. Therefore, each complete student portfolio received 11 or 12 ratings from each of the three or four raters if the portfolio was double rated by the trained raters—four or five ratings for the individual samples and seven ratings for the portfolio.

PILOT STUDY DATA ANALYSES PURPOSES AND EVALUATION FRAMEWORK

The purpose of the pilot data analyses is to evaluate the quality of locally-produced ratings and whether stakes can be associated with the locally-produced ratings. Score reliability is used for evaluating whether stakes can be associated with the pilot study data. For the resulting locally-produced writing score to support high-stakes use, it should have a reliability of at least 0.90 and at least 0.80 or 0.85 for low-stake purposes (Wells & Wollack, 2003[1]). Two potential score-use scenarios are examined.

- ***Can stakes be associated with the locally-produced ratings on any individual writing sample or single portfolio scores?***
  The inter-rater correlation and generalizability coefficient are used for this evaluation. The inter-rater correlation needs to be at least 0.89 for a reliability of at least 0.80. The generalizability coefficient is a rater reliability and needs to be at least 0.80.
- ***Can stakes be associated with the locally-produced writing scores as the sums of seven portfolio rating scores?***
  The generalizability coefficient is used for this evaluation and needs to be at least 0.80.

The year-one pilot data analyses also evaluates the two scoring rubrics that were developed during the school year for quality to inform revision and improvement.

[1] Wells, C. S., & Wollack, J. A. (2003). An instructor's guide to understanding test reliability. Retrieved from http://testing.wisc.edu/Reliability.pdf

- The writing pilot rubric was developed to assign one rating to each individual final writing sample. Raters should be consistently assigning identical or adjacent ratings to the same writing sample. The expectation is that the raters should use all rating categories (i.e., 1, 2, 3, and 4) rather than concentrating on a selected few rating categories. The distributions of ratings will be used for this purpose.
- The writing portfolio rubric was developed to assign seven analytic ratings to each complete student portfolio. Again, the raters should be consistently assigning identical or adjacent ratings to the same writing sample. The correlations among the seven portfolio ratings are used for this evaluation, and intermediate to high correlations among the seven ratings are expected.

Additionally, the relevant statistics of STAAR grades 4 and 7 writing assessments are provided as another frame of reference.

- Spring 2017 STAAR essay scoring has achieved 57% or higher exact agreement and higher than 90% exact or adjacent agreement.
- The reliabilities of the spring 2017 STAAR grades 4 and 7 writing tests were 0.84 and 0.86, respectively. The STAAR English I and English II tests were not included for reference because they measure both reading and writing.

SUMMARY OF STATISTICAL ANALYSIS RESULTS AND CONCLUSIONS

A smaller sample was planned for year one to test the process and rubrics. Higher sample attrition occurred during the implementation. In the end, 36 to 435 for individual writing samples and 153 to 293 for complete student portfolios per grade/course were rated by teachers, ESC raters, and trained raters. Detailed analyses methodology and results are summarized in Appendix C (Rater Scores Summary), D (Rater Correlations and Percentages of Agreement), E (Rater Score Reliability), F (Correlations of Portfolio Scores), and G (Correlations between Writing Pilot Scores and STAAR Summative Writing Scores). The following conclusions are drawn at the end of year-one data analyses. Further generalization of the results should take into consideration the small sample sizes.

No individual or sum of ratings in the current study reached the reliability of 0.80, and most of the scores' reliabilities were far below 0.80.

The correlation and agreement analyses showed that across all four tests, rater scores, and rater pairs, no correlation and exact agreement rate exceeded 0.88 (i.e., the reliabilities were all below 0.80) and 68%, respectively, at the class level for individual ratings. The generalizability coefficients were calculated at the class level as rater reliabilities. Across the four tests and all rating scores (individual ratings and sums of seven portfolio ratings), no rater reliability exceeded 0.65. Employing more than one rater and the adjudication rules used with STAAR writing prompts should increase the score reliability.

The two rubrics worked well, but more training materials and training are warranted.

For all ratings, most of their means were between 2 and 3. It appeared that all four rating categories (i.e., 1, 2, 3, and 4) were used by the raters, which indicates that raters could distinguish the quality of student writings according to the rubrics. The polychoric correlations among the seven portfolio scores within each rater group were, in most cases, intermediate to high (i.e., between 0.45 and 0.90). Therefore, in general, the separation of the seven portfolio scores was justified by the data.

Ratings assigned by trained raters were used as the standard to meet by teachers and ESC raters. When ratings from second trained raters were used to check the quality of ratings assigned by trained raters, their exact agreement rates on the individual writing samples in grade 4 writing and English II were close to or exceeded those rates on the writing prompts in the corresponding spring STAAR tests. For the other scores, the two ETS raters had similar results with the other rater pairs, and most of their exact agreement rates were below those rates on the writing prompts in the corresponding spring STAAR tests. The low consistency of rating scores among the trained raters may be due to scoring rubrics being new and/or lack of sufficient scoring training.

RECOMMENDATIONS

Based on the writing pilot data analysis results from the year-one pilot study, local rating scores cannot support high-stakes use at this time due to their low rater reliabilities. Improvements are recommended that (a) enhance training on the scoring rubrics to improve teachers' rating quality, especially for portfolio scoring; and (b) use more than one teacher as the rater and appropriate adjudication rules to enhance teacher rating reliability.

With the experience gained from the year-one pilot study, processes and systems are being established for the year-two pilot study and potential statewide implementation. Additionally, the sample sizes will be doubled in the year-two pilot study to support the interpretation of results and make recommendations for potential statewide implementation.

Also, the STAAR writing test design can be considered to develop a hybrid model for assessing writing in the future. That is, multiple-choice items can be administered either in the STAAR test or locally in addition to the locally-scored writing samples to increase the reliability of student writing scores for supporting the appropriate stakes.

YEAR-TWO IMPLEMENTATION

For the 2017–2018 school year, the existing partnering regions will expand the number of schools and students participating in the writing pilot to double the sample size from 1,707 students to approximately 3,500 students.

Region 6 (Huntsville) will continue to participate in the writing pilot with two partnering school districts from year one: Calvert ISD and Huntsville ISD. Region 6 will expand participation by adding students from Magnolia ISD and Snook ISD. Region 10 (Richardson) will be expanding participation for year two by adding schools from their current partnering districts: Athens ISD, Garland ISD, and Sunnyvale ISD. Finally, Region 16 (Amarillo) will be participating with a combination of new and returning districts. Dumas ISD will return for year two at their current participation levels. Amarillo ISD will return with all schools from last year and one additional high school for year two. To expand participation, Region 16 will add new partnering districts: Memphis ISD, LeFors ISD, and Kress ISD. The total number of districts participating in year two will expand from seven to twelve school districts.

Year two of the pilot study will include the collection and scoring of a range of student writing samples produced throughout the school year. The writing products to be completed, submitted, and scored for the 2017–2018 school year are:

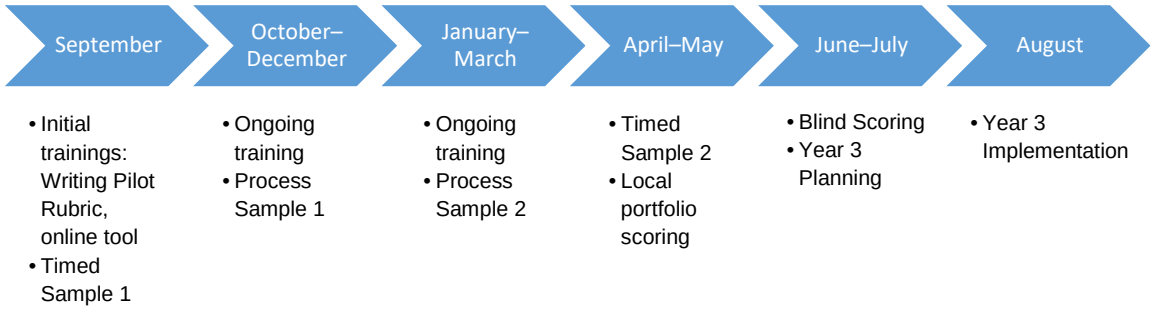
- two timed writing samples completed at the beginning and end of the school year based on a specific writing prompt chosen by each student from a selection of three prompts;
- two instructional writing process samples from different genres (i.e., personal narrative, expository, persuasive, or analytic) that include evidence of a writing process from start to finish (e.g., planning, drafting, revising, editing, and publishing); and
- an instructional portfolio containing the writing samples listed above.

To assist with the collection of samples and scoring, participating districts will be using a new online collection tool developed specifically for the writing pilot. The new tool will give teachers and students the option to type their timed samples directly into the online platform or upload samples as attachments. In addition to housing all the required student samples, the online platform will give teachers the opportunity to view their students' work alongside the writing pilot rubrics, therefore making scoring of the samples for year two much more streamlined and efficient. The 2017–2018 school year will serve as the beta year for the new online tool with the plan to expand the tool functions and capabilities for the 2018–2019 school year.

The scoring for year two will be conducted the same way as year one, consisting of several components. The student samples will initially be scored by each student's teacher-of-record. Additionally, the samples will receive a blind score that will be coordinated at the local level by each participating ESC and include local teachers who are certified to teach English language arts. A final sampling of scores will be conducted by TEA and ETS.

Year two will include improved and expanded training on the Writing Pilot Rubric and the Portfolio Rubric. In addition, districts will receive training on the new online collection tool aimed at streamlining the collection and scoring of writing pilot samples.

Year-two overview timeline:



SUMMARY

Ultimately, a well-designed assessment should inform and aid best practices in instruction. Therefore, the goal of the writing pilot is to support the growth of Texas students as effective writers. Through the hard work and dedication of all participants to date, TEA is extremely positive about the progress achieved during year one and the upcoming accomplishments and possibilities going into year two of the writing pilot.

Appendix A: Writing Pilot Rubric

Texas Writing Pilot Program Rubric 2016–2017

Score Point 4 (Accomplished): The response will contain most of the following characteristics.			
Organizational Structure and Focus	**Content/Development of Ideas**	**Use of Language**	**Conventions**
• Structure is clearly appropriate to the purpose. • The writer establishes and maintains a strong focus. • Strong, meaningful transitions and idea-to-idea, sentence-to-sentence, and paragraph-to-paragraph connections are clearly evident.	• Specific, well chosen, and relevant details are clearly evident. • Ideas are clearly, thoughtfully, and effectively expressed and developed.	• Language and word choice are purposeful, precise, and enhance the writing. • Sentences are purposeful, well-constructed, and controlled. • Use of an authentic, expressive voice is clearly reflected throughout the writing.	• Although minor errors may be evident, they do not detract from the fluency or clarity of the writing. • Use of grade-appropriate spelling, capitalization, punctuation, grammar, and usage conventions is consistently demonstrated.
Score Point 3 (Satisfactory): The response will contain most of the following characteristics.			
Organizational Structure and Focus	**Content/Development of Ideas**	**Use of Language**	**Conventions**
• Structure is, for the most part, appropriate to the purpose. • The writer, for the most part, establishes and maintains focus. • Sufficient use of transitions and idea-to-idea, sentence-to-sentence, and paragraph-to-paragraph connections is somewhat evident.	• Specific, appropriate, and relevant details are somewhat evident. • Ideas are sufficiently expressed and developed.	• Language and word choice are, for the most part, clear, concise, and somewhat enhance the writing. • Sentences are somewhat purposeful and adequately constructed and controlled. • Authentic voice is somewhat evident and appropriately reflected throughout the writing.	• Minor errors create some disruption in the fluency or clarity of the writing. • Use of grade-appropriate spelling, capitalization, punctuation, grammar, and usage conventions is adequately demonstrated.

Texas Education Agency

Score Point 2 (Basic): The response will contain most of the following characteristics.			
Organizational Structure and Focus	**Content/Development of Ideas**	**Use of Language**	**Conventions**
• Structure is evident but may not always be appropriate to the purpose. • The writer does not effectively establish or maintain focus and may include irrelevant information. • Use of transitions, idea-to-idea, sentence-to-sentence, and paragraph-to-paragraph connections is minimal or inconsistent.	• Specific and relevant details are too brief, too vague, or are not clearly evident. • Ideas are minimally expressed and developed.	• Language and word choice are general, imprecise, or inappropriate and do not sufficiently enhance the writing. • Sentences are awkward or only somewhat controlled. • Authentic voice is inconsistent throughout the writing.	• Distracting errors create moderate disruptions in the fluency or clarity of the writing. • Use of grade-appropriate spelling, capitalization, punctuation, grammar, and usage conventions is partially demonstrated.
Score Point 1 (Very Limited): The response will contain most of the following characteristics.			
Organizational Structure and Focus	**Content/Development of Ideas**	**Use of Language**	**Conventions**
• Structure is inappropriate to the purpose. • Focus is not established or maintained. • Transitions, idea-to-idea, sentence-to-sentence, and paragraph-to-paragraph connections are not evident.	• Details are inappropriate or missing. • Ideas are missing or not expressed or developed.	• Language and word choice is limited or missing and does not enhance the writing. • Sentences are simplistic or uncontrolled. • Authentic voice is missing or inappropriate to the writing task.	• Serious and persistent errors create disruptions in the fluency or clarity of the writing. • Little to no use of grade-appropriate spelling, capitalization, punctuation, grammar, and usage conventions is demonstrated.

Texas Writing Pilot Rubric 2016–2017

Appendix B: Portfolio Rubric

Texas Writing Pilot End-of-Year Portfolio Rubric

Directions: This rubric is used to evaluate a student's overall portfolio from the 2017—2018 Texas Writing Pilot. When reviewing the portfolio, consider all items as evidence of the student's engagement in the writing process from the start of the school year through the end and as evidence of their overall development as a writer. As you are analyzing the artifacts, follow the score point descriptors below to assign a score for **each** row.

Score Point Descriptions

Score Point 4—Exceeds Standards:	Score Point 3—Meets Standards:	Score Point 2—Developing Towards Standards:	Score Point 1—Substantially Below Standards:
From the artifacts, there is **clear and compelling evidence** that the student met most of the expectations.	From the artifacts, there is **satisfactory evidence** that the student met most of the expectations.	From the artifacts, there is **limited or inconsistent evidence** that the student met most of the expectations.	From the artifacts, there is **little to no evidence** that the student met most of the expectations.

Portfolio Categories and Expectations	Evidence and Examples	Score
Planning: A range of strategies is used to generate ideas for writing that are appropriate to the task, topic, or genre.	Evidence demonstrates that the student has planned for writing by generating ideas relevant to the task, topic, or genre. Evidence **could** include webs, graphic organizers, journals, drawings, brainstorms, outlines, reflection(s) from conversation(s) with teacher(s) or classroom discussion(s), etc.	
Drafting: Ideas from planning activities are categorized, organized, and/or developed into drafts that are appropriate to the task, topic, or genre.	Evidence demonstrates that ideas from planning are developed to make the writing more coherent and that the student has taken his/her ideas and turned them into a written format to convey thoughts that are appropriate to the task, topic, or genre. Evidence **could** include expanded outlines, expanded webs, expanded graphic organizers, paragraphs, the first draft, etc.	
Revising: Drafts are revised for coherence, organization, use of language, sentence structure, transitions, connections, and/or to add/delete ideas that clarify for meaning as appropriate to the task, topic, or genre.	Evidence demonstrates that ideas from first draft(s) are further developed for coherence, organization, sentence structure, transitions, and/or use of language as appropriate to the task, topic, or genre. Evidence **could** include drafts, reflection(s) from conversation(s) with teacher(s), peer editing, or classroom discussion(s), etc.	
Editing, Publishing and Attention to Feedback: Drafts are edited for appropriate grammar, mechanics, and/or spelling as appropriate to the task, topic, or genre and feedback from peers and/or teachers is incorporated to create texts ready for publication as appropriate to the task, topic, or genre.	Evidence demonstrates that revisions have been made and/or suggestions from self, peer(s), and/or teacher(s) have been incorporated. Evidence **could** include copy/copies of a full piece of writing with mark-up for grammar, mechanics, and/or spelling that are appropriate to the task, topic, or genre and/or a final, clean, publication ready copy of the written piece appropriate to the task, topic, or genre.	
Expressing Ideas: The student effectively communicates and expresses himself/herself through writing in a variety of genres, including the following; Personal Narrative, Expository, Persuasive, and/or Analytical, as appropriate to grade level.	Evidence demonstrates that the writing has specific, well chosen, and relevant details that are clearly, thoughtfully, and effectively expressed and developed as appropriate to the genre and grade level.	
Organization and Structure: The student effectively expresses his/her ideas through composed texts that are organized, structured, and focused as appropriate to grade level.	Evidence demonstrates a strong focus, meaningful transitions, and idea-to-idea, sentence-to-sentence, and paragraph-to-paragraph connections.	
Use of Language and Conventions: The student expresses his/her ideas through composed texts that effectively exhibit use of language, word choice, sentence structure, voice, and conventions that are appropriate to the genre and grade level.	Evidence demonstrates language and word choice that are purposeful, precise, and enhance the writing; purposeful, well-constructed, and controlled sentences; use of an authentic, expressive voice; and use of grade-appropriate spelling, capitalization, punctuation, grammar, and usage conventions.	

Appendix C: Rater Scores Summary

Tables C1–C4 show the summary statistics for the individual writing sample scores and portfolio scores across all students for grade 4 writing, grade 7 writing, English I, and English II, respectively: number of responses (N), score mean (Mean), standard deviation (SD), and percentage of students in each rating category (S1, S2, S3, S4). Because there is no prompt indicator other than TS1, PS1, PS2, PS3, and TS2 in the dataset, we assumed common writing prompts within a class but different ones across classes. Note that teachers chose and will continue to choose writing prompts for their classes.

Teachers assigned ratings of 36 to 435 for individual writing samples and 153 to 293 for complete student portfolios. One noteworthy observation is that for PS3 and TS2 in English I and English II as well as PS1 in English II, teachers missed a lot of rating scores as the number of rating scores given by the ESC raters (ESC) and the first TEA trained raters (TEA1) were much higher than those by teachers. For all ratings, most of their means were between 2 and 3. It appeared that all four rating categories (i.e., 1, 2, 3, and 4) were used by the raters, which indicates that raters were able to distinguish the quality of student writings according to the rubrics.

Table C1. Rater Scores Summary: Grade 4 Writing

Score	Rater: Teacher							Rater: ESC							Rater: TEA1							Rater: TEA2						
	N	Mean	SD	S1 (%)	S2 (%)	S3 (%)	S4 (%)	N	Mean	SD	S1 (%)	S2 (%)	S3 (%)	S4 (%)	N	Mean	SD	S1 (%)	S2 (%)	S3 (%)	S4 (%)	N	Mean	SD	S1 (%)	S2 (%)	S3 (%)	S4 (%)
TS1	435	1.63	0.72	51	36	13	0	306	1.69	0.72	46	40	14	0	305	1.66	0.69	47	41	12	0	60	1.55	0.62	52	42	7	0
PS1	412	2.09	0.90	30	38	25	7	305	2.04	0.85	28	46	20	6	309	1.87	0.74	32	51	14	3	59	1.80	0.76	37	49	10	3
PS2	331	2.20	0.89	24	41	27	8	189	2.04	0.88	32	38	25	5	197	1.87	0.83	37	44	15	5	54	1.80	0.86	44	35	17	4
PS3	196	2.69	0.92	10	33	36	21	111	2.51	0.87	10	44	31	15	111	2.29	0.91	20	42	27	11	0						
TS2	285	2.47	0.96	18	31	36	15	296	2.06	0.80	24	50	21	5	305	1.91	0.78	32	47	18	3	60	1.90	0.75	30	53	13	3
PF1	293	2.96	0.76	4	19	54	23	307	2.85	0.99	10	26	31	33	307	2.82	0.86	9	21	50	21	49	2.84	0.80	8	16	59	16
PF2	293	2.92	0.73	4	17	60	18	307	2.75	0.91	9	30	37	23	307	2.71	0.88	11	25	47	17	49	2.82	0.73	6	18	63	12
PF3	293	2.41	0.82	13	41	38	8	300	2.24	0.94	25	35	30	10	307	2.22	0.82	20	43	32	5	49	2.45	0.87	14	37	39	10
PF4	293	2.61	0.79	9	33	48	11	296	2.36	0.95	21	33	34	12	307	2.27	0.80	15	49	29	7	49	2.59	0.86	8	41	35	16
PF5	293	2.62	0.95	14	30	37	19	306	2.32	0.88	19	37	35	8	306	2.46	0.80	10	43	37	9	49	2.51	0.79	10	37	45	8
PF6	293	2.50	0.92	16	31	39	14	307	2.30	0.87	19	40	33	8	307	2.22	0.79	18	46	32	4	49	2.41	0.81	12	43	37	8
PF7	293	2.46	0.93	17	33	36	14	306	2.22	0.90	24	40	28	8	307	2.15	0.78	20	50	27	4	49	2.35	0.86	14	47	29	10

Note.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

SD=Standard Deviation.

S1, S2, S3, S4=Percentage of students receiving Scores 1 to 4, respectively.

Table C2. Rater Scores Summary: Grade 7 Writing

Score	Rater: Teacher							Rater: ESC							Rater: TEA1							Rater: TEA2						
	N	Mean	SD	S1 (%)	S2 (%)	S3 (%)	S4 (%)	N	Mean	SD	S1 (%)	S2 (%)	S3 (%)	S4 (%)	N	Mean	SD	S1 (%)	S2 (%)	S3 (%)	S4 (%)	N	Mean	SD	S1 (%)	S2 (%)	S3 (%)	S4 (%)
TS1	313	2.19	0.82	21	44	30	5	204	2.12	0.83	22	51	20	7	209	2.08	0.86	29	39	28	4	36	1.92	0.84	36	39	22	3
PS1	309	2.60	0.94	13	32	36	19	207	2.48	1.00	16	42	21	22	210	2.40	0.94	19	37	31	14	37	2.54	0.96	14	38	30	19
PS2	303	2.58	0.86	10	36	40	14	68	2.06	0.83	25	50	19	6	71	2.13	0.91	27	42	23	8	0						
PS3	141	2.98	0.68	2	18	60	20	139	2.88	0.80	3	30	43	24	139	2.99	0.88	4	26	36	34	0						
TS2	300	2.73	0.77	5	31	50	14	207	2.38	0.87	15	43	31	11	209	2.47	0.89	15	36	37	12	9	1.78	0.67	33	56	11	0
PF1	208	3.17	0.89	5	17	33	45	207	2.90	0.98	10	24	33	33	210	3.13	0.79	3	17	45	36	27	3.00	0.96	7	22	33	37
PF2	208	2.86	0.85	5	30	40	25	206	2.94	0.87	6	24	41	29	210	3.10	0.78	3	17	47	33	36	2.47	1.03	17	42	19	22
PF3	208	2.67	0.91	11	29	41	19	204	2.45	0.91	13	45	26	16	210	2.95	0.84	6	20	47	27	48	2.83	0.83	4	31	42	23
PF4	208	2.70	0.79	2	44	36	18	206	2.35	0.94	19	39	28	13	209	2.97	0.84	6	20	46	28	48	2.83	0.78	4	27	50	19
PF5	208	2.95	0.80	2	27	43	27	207	2.58	0.93	10	42	28	20	210	2.83	0.76	4	27	51	18	48	2.92	0.77	2	27	48	23
PF6	208	2.89	0.75	1	29	48	22	207	2.52	0.89	11	42	31	16	210	2.78	0.76	3	34	46	18	48	2.81	0.82	4	31	44	21
PF7	207	2.79	0.73	2	32	50	15	207	2.39	0.88	14	45	28	13	210	2.72	0.79	6	30	48	15	48	2.77	0.75	4	29	52	15

Note.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

SD=Standard Deviation.

S1, S2, S3, S4=Percentage of students receiving Scores 1 to 4, respectively.

Table C3. Rater Scores Summary: English I

Score	Rater: Teacher							Rater: ESC							Rater: TEA1							Rater: TEA2						
	N	Mean	SD	S1 (%)	S2 (%)	S3 (%)	S4 (%)	N	Mean	SD	S1 (%)	S2 (%)	S3 (%)	S4 (%)	N	Mean	SD	S1 (%)	S2 (%)	S3 (%)	S4 (%)	N	Mean	SD	S1 (%)	S2 (%)	S3 (%)	S4 (%)
TS1	332	1.93	0.80	33	45	20	3	153	2.07	0.83	25	49	20	6	151	2.13	0.96	30	38	22	11	42	2.10	0.79	24	45	29	2
PS1	152	2.01	0.92	36	33	25	6	153	2.02	0.82	28	46	21	5	152	2.11	0.95	29	41	19	11	42	2.05	0.85	29	43	24	5
PS2	36	2.92	0.77	3	25	50	22	20	2.55	0.89	10	40	35	15	20	2.65	0.88	10	30	45	15	0						
PS3	99	2.07	0.98	36	28	27	8	133	2.20	1.00	29	35	23	13	130	2.17	0.94	28	37	26	9	41	2.05	0.89	29	44	20	7
TS2	130	2.08	0.85	28	41	27	5	153	2.03	0.88	31	42	20	7	152	2.05	0.89	30	43	20	7	43	1.84	0.87	42	37	16	5
PF1	153	2.46	0.71	10	37	50	3	153	2.60	0.93	11	37	32	20	151	2.85	0.90	5	32	34	28	48	2.69	0.93	6	44	25	25
PF2	153	2.54	0.61	2	46	48	4	152	2.68	0.90	11	28	43	18	151	2.63	0.84	9	34	42	15	48	2.44	0.82	10	46	33	10
PF3	153	2.24	0.66	11	56	32	1	151	2.08	0.94	32	37	23	9	151	2.11	0.87	28	37	30	5	48	2.04	0.87	27	50	15	8
PF4	153	2.19	0.59	8	67	23	2	152	2.02	0.90	33	39	22	7	151	2.25	0.89	23	38	31	8	48	2.17	1.02	29	40	17	15
PF5	153	2.38	0.68	7	54	35	5	152	2.30	0.75	13	51	32	5	151	2.41	0.87	14	42	32	11	48	2.25	0.81	17	48	29	6
PF6	153	2.24	0.69	9	63	22	5	153	2.40	0.76	12	42	41	5	151	2.37	0.87	17	37	37	9	48	2.31	0.88	17	46	27	10
PF7	153	2.24	0.61	7	63	27	2	153	2.27	0.76	14	48	33	5	151	2.26	0.84	18	46	28	8	48	2.15	0.99	31	33	25	10

Note.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

SD=Standard Deviation.

S1, S2, S3, S4=Percentage of students receiving Scores 1 to 4, respectively.

Table C4. Rater Scores Summary: English II

Score	Rater: Teacher							Rater: ESC							Rater: TEA1							Rater: TEA2						
	N	Mean	SD	S1 (%)	S2 (%)	S3 (%)	S4 (%)	N	Mean	SD	S1 (%)	S2 (%)	S3 (%)	S4 (%)	N	Mean	SD	S1 (%)	S2 (%)	S3 (%)	S4 (%)	N	Mean	SD	S1 (%)	S2 (%)	S3 (%)	S4 (%)
TS1	423	1.93	0.74	29	51	18	2	241	2.08	0.80	25	45	27	3	240	2.15	0.84	23	44	28	5	56	1.93	0.81	30	52	13	5
PS1	156	1.92	0.69	28	52	20	0	242	2.33	0.87	17	41	32	9	242	2.25	0.88	18	50	21	11	55	2.15	0.89	24	47	20	9
PS2	55	1.98	0.68	24	55	22	0	37	2.22	0.71	14	54	30	3	37	2.05	0.66	19	57	24	0	22	2.09	0.68	14	68	14	5
PS3	98	2.11	0.88	32	28	39	2	202	2.08	0.79	24	48	25	3	201	2.12	0.85	25	44	25	6	33	2.12	0.89	27	39	27	6
TS2	149	2.07	0.81	24	50	20	5	242	2.13	0.76	19	51	26	3	242	2.16	0.83	24	41	31	4	56	2.05	0.80	27	43	29	2
PF1	238	2.62	0.86	11	32	44	14	240	2.58	0.97	14	34	31	20	241	2.54	0.84	9	41	37	13	66	2.71	0.80	6	32	47	15
PF2	238	2.73	0.69	3	33	53	11	241	2.88	0.81	4	28	45	23	241	2.62	0.79	6	39	42	13	66	2.77	0.80	6	27	50	17
PF3	238	2.22	0.85	23	37	36	5	241	2.35	0.87	17	41	32	10	241	2.36	0.82	14	44	34	8	66	2.64	0.78	9	27	55	9
PF4	238	2.29	0.90	23	32	38	7	239	2.29	0.86	17	48	26	10	241	2.43	0.80	11	44	36	9	66	2.56	0.86	12	32	44	12
PF5	238	2.63	0.75	6	36	48	11	240	2.33	0.80	12	52	27	9	241	2.43	0.76	9	46	37	7	66	2.59	0.72	5	41	45	9
PF6	238	2.56	0.86	11	34	42	13	241	2.44	0.78	10	43	39	7	241	2.34	0.79	12	49	32	7	66	2.48	0.81	11	39	41	9
PF7	238	2.59	0.86	12	30	45	13	240	2.28	0.73	11	56	28	6	241	2.23	0.78	16	49	29	5	66	2.33	0.79	12	50	30	8

Note.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

SD=Standard Deviation.

S1, S2, S3, S4=Percentage of students receiving Scores 1 to 4, respectively.

Appendix D: Rater Correlations and Percentages of Agreement

The correlations included in Tables D1–D8 are polychoric correlations. Polychoric correlation is suitable for the case where both variables are ordered categorical variables (Drasgow, 1988[2]), like rating scores in this study. Polychoric correlation assumes there is a continuous variable underlying each categorical variable and the two continuous variables follow a binormal distribution. The polychoric correlation is the correlation between the two variables in the binormal distribution. Polychoric correlation is estimated by the maximum likelihood estimation. Compared to Pearson correlation, polychoric correlation more accurately reflects the true relationship between two ordered categorical variables if the assumptions hold, while Pearson correlation tends to underestimate the association.

Tables D1–D4 list the sample sizes, polychoric correlations (Cor), percentages of exact agreement (EA), and percentages of exact or adjacent agreement (EAA) among rating scores from the three raters—teacher, ESC rater, and TEA rater 1—for each of the 12 rating scores in the four tests, respectively. The correlations in Tables D1–D4 are plotted in Figures B1–B4 and the percentages of exact agreement in Tables D1–D4 are plotted in Figures D5–D8 for visual observation. For each calculation of correlation and agreement rates, the sample size needed to be at least 30.

Across the four tests and rater pairs:

- the mean correlations over the 12 rating scores were between 0.37 and 0.58;
- the mean percentages of exact agreement over the 12 rating scores ranged from 39% to 47%;
- the mean percentages of exact or adjacent agreement over the 12 rating scores ranged from 87% to 94%; and
- the maximum correlation, exact agreement rate, and exact or adjacent agreement rate across the 12 rating scores were 0.69, 61%, and 100%, respectively.

There was not a general pattern that teachers' rating quality improved across the school year by these measures; that is, these statistics for TS2 are not necessarily better than for TS1, for example. As a framework to assist interpretation, the exact agreement rates for the writing prompts in the spring 2017 STAAR paper administration were 58%, 59%, 58%, and 57%, respectively, in grade 4, grade 7, English I, and English II.

Tables D1–D4 also compare the correlations and agreement rates between two trained TEA raters (TEA1 vs. TEA2) as a quality check for the trained raters. For grade 4 and English II, two trained TEA raters had the most consistent rating scores on the writing sample scores compared to the other rater pairs: the correlations ranged from 0.58 to 0.92; the exact agreement rates ranged from 51% to 73%; and the exact or adjacent rates ranged from 93% to 100%. For the other scores, the two trained TEA raters had similar correlations and agreement rates as the other pairs of raters. It's worth noting that the first four portfolio scores (i.e., Planning,

[2] Drasgow, F. (1988). Polychoric and polyserial correlations. In L. Kotz, & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*. Vol. 7 (pp. 69-74). New York: Wiley.

Drafting, Revising, Editing, Publishing and Attention) in grade 4 had the lowest correlations and agreement rates across all pairs of raters, and the Planning score in English I had a correlation of almost 0 between teacher and ESC or TEA1 rater.

The polychoric correlations and agreement rates were also calculated among teacher, ESC rater and TEA rater 1 at the class level and summarize the results across classes in Tables D5–D8 for the four tests, respectively. Two timed writing samples were stacked as the timed writing sample and the three process writing samples were stacked as the process writing sample so as to increase the sample size of each writing type in a class. There were some variations in the correlations and agreement rates at the class level in all four tests. The maximum correlation and exact agreement rate for a class across all subjects, rater pairs, and rating scores were 0.88 and 68%, respectively.

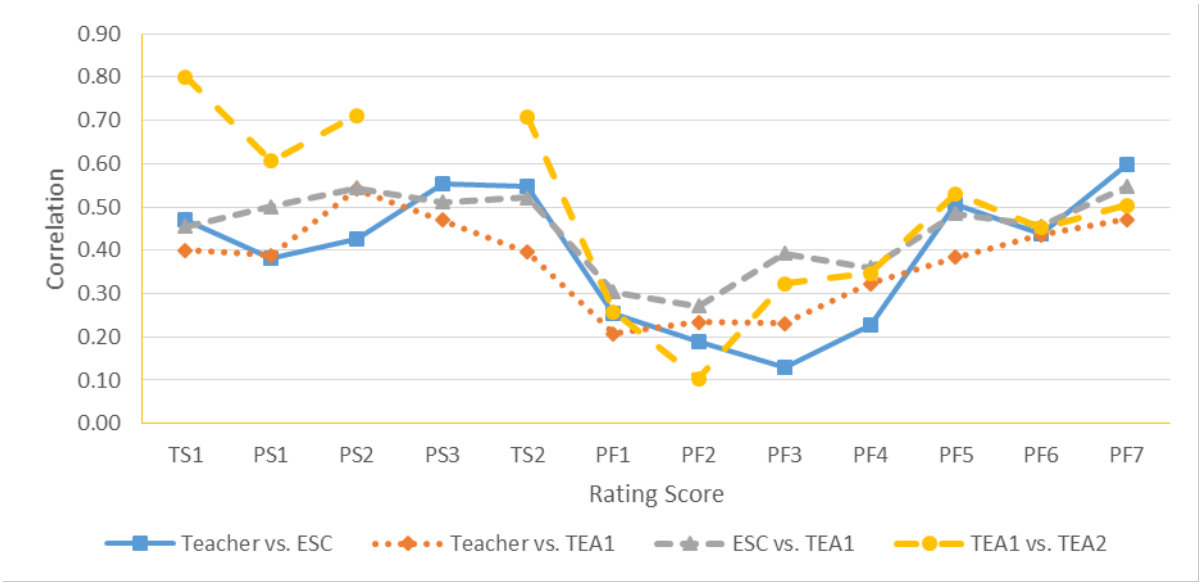


Figure D1. Rater correlations on grade 4 writing rating scores.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

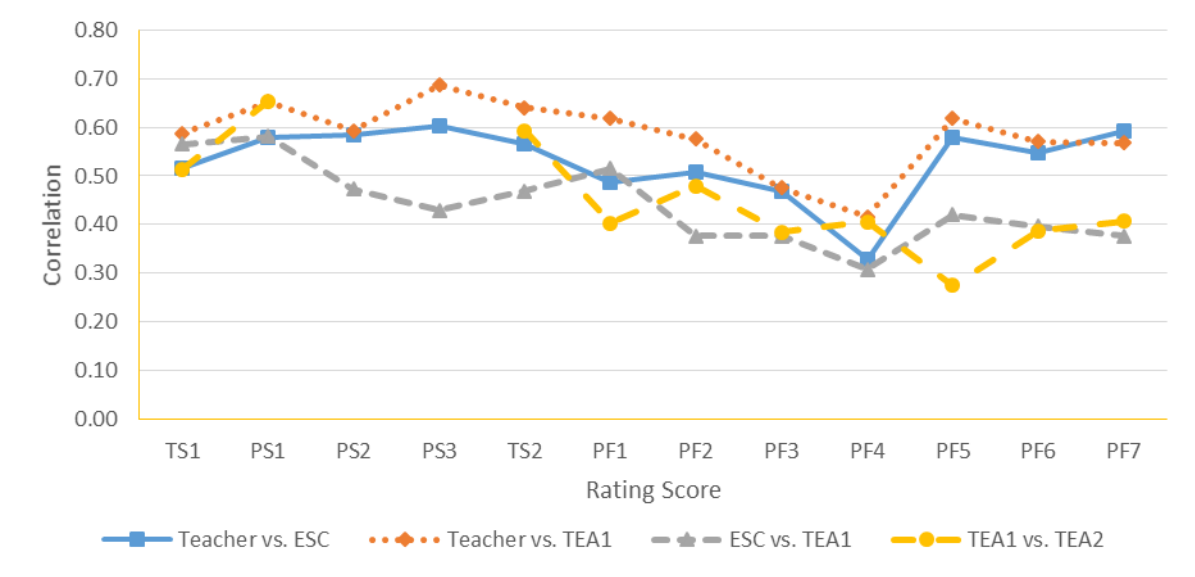


Figure D2. Rater correlations on grade 7 writing rating scores.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

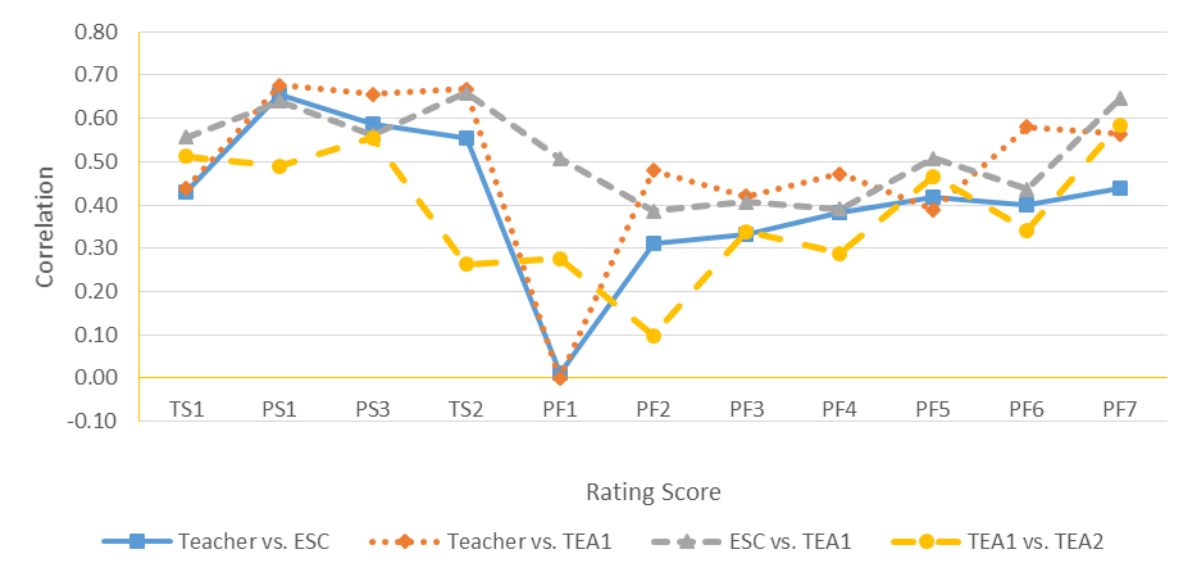


Figure D3. Rater correlations on English I rating scores.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

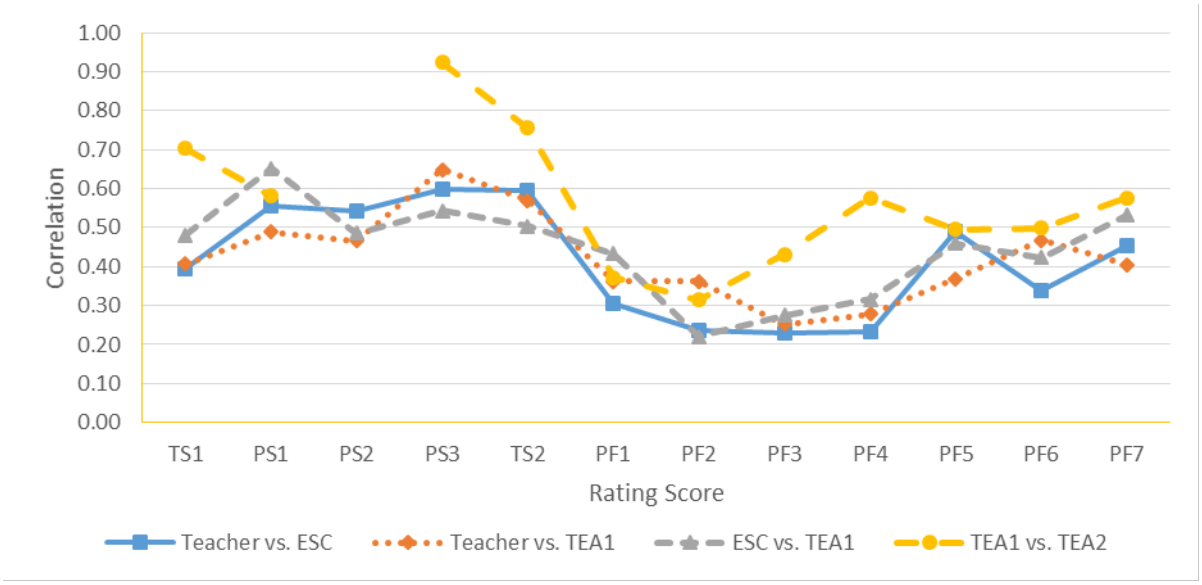


Figure D4. Rater correlations on English II rating scores.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

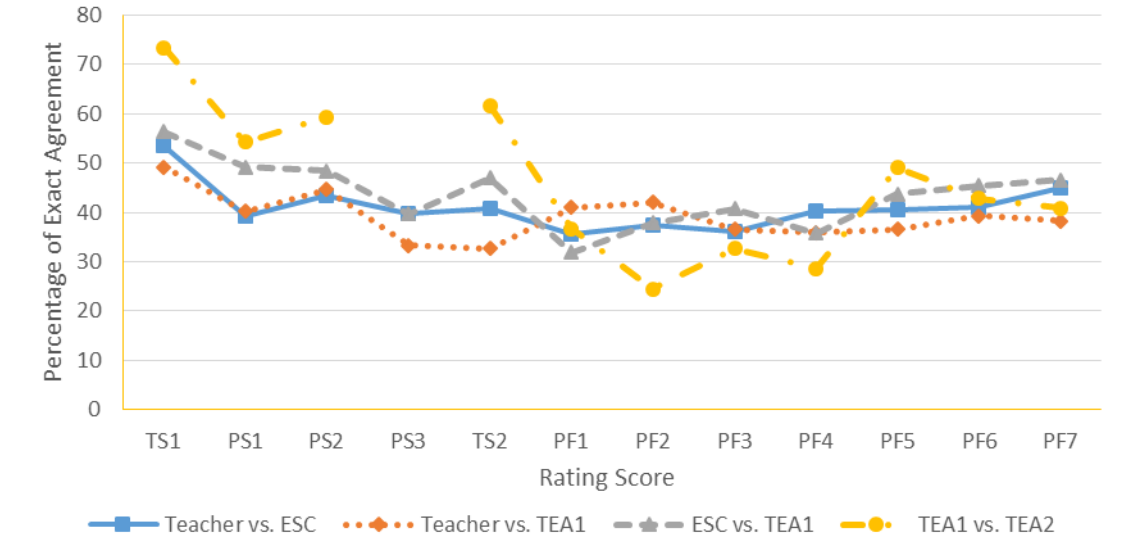


Figure D5. Percentage of exact rater agreement on grade 4 writing rating scores.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

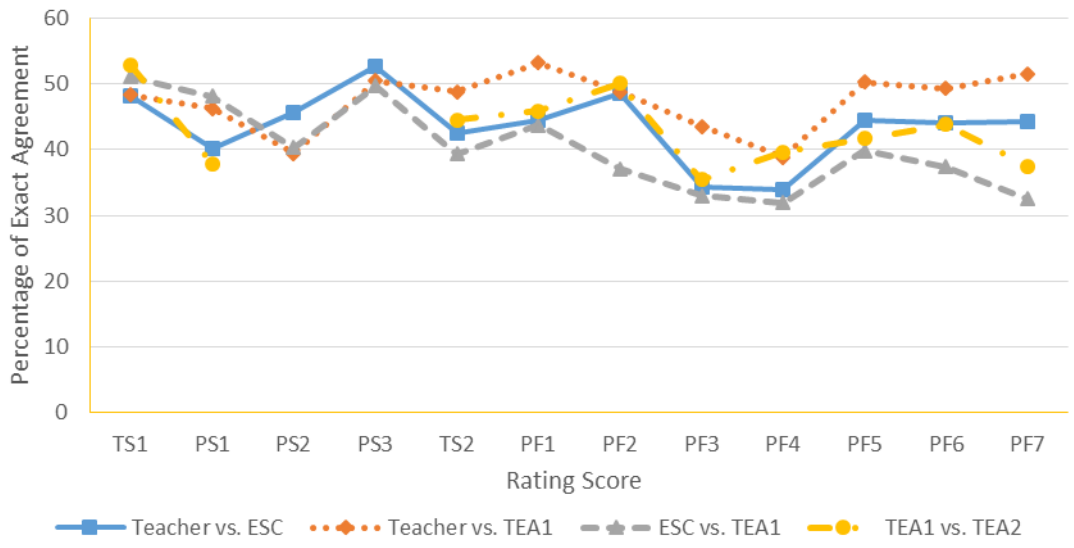


Figure D6. Percentage of exact rater agreement on grade 7 writing rating scores.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

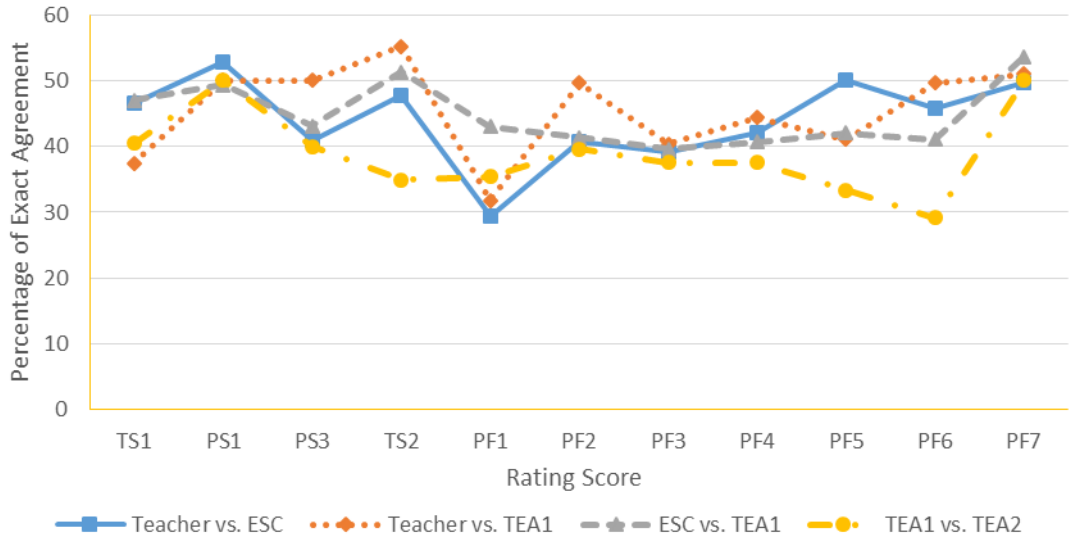


Figure D7. Percentage of exact rater agreement on English I rating scores.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

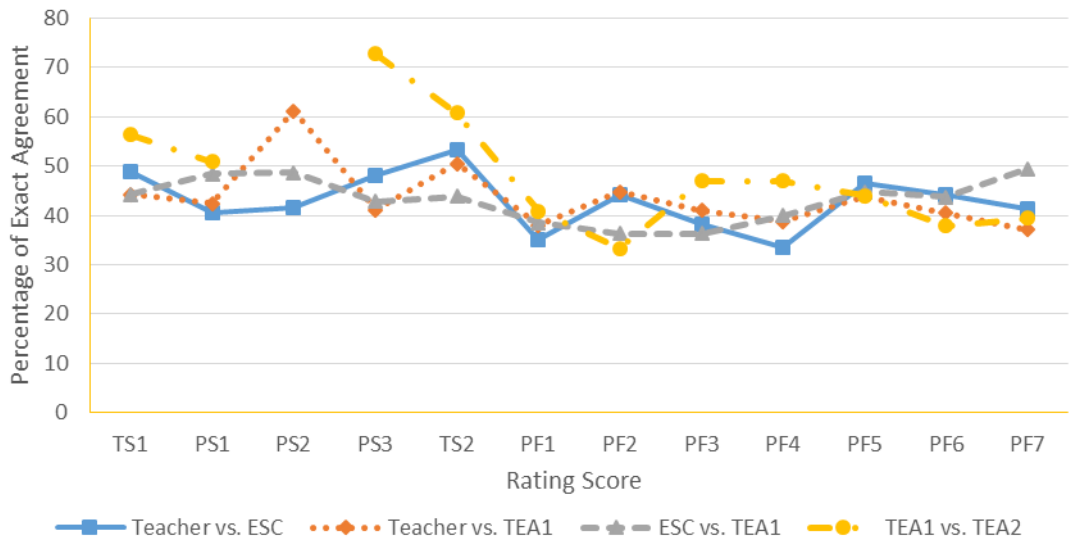


Figure D8. Percentage of exact rater agreement on English II rating scores.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

Table D1. Rater Correlations and Percentages of Agreement: Grade 4 Writing

Score	Teacher vs. ESC				Teacher vs. TEA1				ESC vs. TEA1				TEA1 vs. TEA2			
	N	Cor	EA (%)	EAA (%)	N	Cor	EA (%)	EAA (%)	N	Cor	EA (%)	EAA (%)	N	Cor	EA (%)	EAA (%)
TS1	306	0.47	54	93	305	0.40	49	93	302	0.45	56	93	60	0.80	73	100
PS1	304	0.38	39	88	308	0.39	40	88	303	0.50	49	92	59	0.61	54	97
PS2	113	0.43	43	84	121	0.54	45	91	188	0.54	48	91	54	0.71	59	93
PS3	111	0.55	40	94	111	0.47	33	86	111	0.51	40	89				
TS2	204	0.55	41	90	211	0.39	33	83	294	0.52	47	94	60	0.71	62	97
PF1	293	0.25	35	86	290	0.21	41	86	304	0.30	32	83	49	0.26	37	86
PF2	292	0.19	37	85	290	0.23	42	84	304	0.27	38	82	49	0.10	24	84
PF3	285	0.13	36	80	290	0.23	37	87	297	0.39	41	89	49	0.32	33	88
PF4	281	0.23	40	83	290	0.32	36	89	293	0.36	36	87	49	0.35	29	88
PF5	291	0.51	41	87	290	0.38	37	87	302	0.48	44	93	49	0.53	49	94
PF6	292	0.44	41	88	290	0.43	39	87	304	0.46	45	91	49	0.45	43	90
PF7	291	0.60	45	91	290	0.47	38	89	303	0.55	47	93	49	0.50	41	94
Mean	255	0.39	41	87	257	0.37	39	87	275	0.45	44	90	52	0.49	46	92
SD	72	0.16	5	4	70	0.11	5	3	61	0.09	7	4	5	0.21	15	5
Max	306	0.60	54	94	308	0.54	49	93	304	0.55	56	94	60	0.80	73	100
Min	111	0.13	35	80	111	0.21	33	83	111	0.27	32	82	49	0.10	24	84

Note.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

SD=Standard Deviation, Cor=Correlation, EA=Percentage of exact agreement, EAA=Percentage of exact or adjacent agreement.

Table D2. Rater Correlations and Percentages of Agreement: Grade 7 Writing

Score	Teacher vs. ESC				Teacher vs. TEA1				ESC vs. TEA1				TEA1 vs. TEA2			
	N	Cor	EA (%)	EAA (%)	N	Cor	EA (%)	EAA (%)	N	Cor	EA (%)	EAA (%)	N	Cor	EA (%)	EAA (%)
TS1	204	0.52	48	94	209	0.59	48	96	204	0.56	51	93	36	0.51	53	89
PS1	207	0.58	40	90	210	0.65	46	92	206	0.58	48	87	37	0.65	38	95
PS2	68	0.58	46	93	71	0.59	39	94	67	0.47	40	91				
PS3	131	0.60	53	97	131	0.69	50	98	139	0.43	50	88				
TS2	207	0.57	43	92	209	0.64	49	94	206	0.47	39	89	36	0.59	44	86
PF1	207	0.49	44	89	207	0.62	53	94	206	0.51	44	90	48	0.40	46	94
PF2	206	0.51	49	90	207	0.58	49	93	205	0.38	37	91	48	0.48	50	98
PF3	204	0.47	34	89	207	0.47	43	89	203	0.38	33	84	48	0.38	35	92
PF4	206	0.33	34	86	206	0.42	39	89	204	0.31	32	79	48	0.41	40	90
PF5	207	0.58	44	88	207	0.62	50	97	206	0.42	40	87	48	0.28	42	92
PF6	207	0.55	44	91	207	0.57	49	98	206	0.40	37	88	48	0.39	44	92
PF7	206	0.59	44	92	206	0.57	51	96	206	0.38	33	89	48	0.41	38	96
Mean	188	0.53	44	91	190	0.58	47	94	188	0.44	40	88	45	0.45	43	92
SD	44	0.08	5	3	43	0.07	5	3	43	0.08	7	3	6	0.11	6	3
Max	207	0.60	53	97	210	0.69	53	98	206	0.58	51	93	48	0.65	53	98
Min	68	0.33	34	86	71	0.42	39	89	67	0.31	32	79	36	0.28	35	86

Note.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

SD=Standard Deviation, Cor=Correlation, EA=Percentage of exact agreement, EAA=Percentage of exact or adjacent agreement.

Table D3. Rater Correlations and Percentages of Agreement: English I

Score	Teacher vs. ESC				Teacher vs. TEA1				ESC vs. TEA1				TEA1 vs. TEA2			
	N	Cor	EA (%)	EAA (%)	N	Cor	EA (%)	EAA (%)	N	Cor	EA (%)	EAA (%)	N	Cor	EA (%)	EAA (%)
TS1	133	0.43	47	90	131	0.44	37	89	151	0.56	47	91	42	0.51	40	93
PS1	91	0.65	53	93	90	0.68	50	91	152	0.64	49	93	42	0.49	50	93
PS2																
PS3	71	0.59	41	89	70	0.66	50	90	130	0.56	43	91	40	0.56	40	93
TS2	88	0.55	48	92	87	0.67	55	93	152	0.66	51	95	43	0.26	35	86
PF1	153	0.01	29	81	151	0.00	32	76	151	0.51	43	87	48	0.28	35	90
PF2	152	0.31	41	89	151	0.48	50	94	150	0.39	41	87	48	0.10	40	88
PF3	151	0.33	39	88	151	0.42	40	93	149	0.41	40	87	48	0.34	38	85
PF4	152	0.38	42	91	151	0.47	44	94	150	0.39	41	87	48	0.29	38	79
PF5	152	0.42	50	95	151	0.39	41	91	150	0.51	42	94	48	0.47	33	96
PF6	153	0.40	46	93	151	0.58	50	94	151	0.44	41	92	48	0.34	29	94
PF7	153	0.44	50	96	151	0.56	51	96	151	0.65	54	96	48	0.58	50	92
Mean	132	0.41	44	91	130	0.49	46	91	149	0.52	45	91	46	0.38	39	90
SD	32	0.17	7	4	32	0.19	7	5	6	0.10	5	3	3	0.15	6	5
Max	153	0.65	53	96	151	0.68	55	96	152	0.66	54	96	48	0.58	50	96
Min	71	0.01	29	81	70	0.00	32	76	130	0.39	40	87	40	0.10	29	79

Note.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

SD=Standard Deviation, Cor=Correlation, EA=Percentage of exact agreement, EAA=Percentage of exact or adjacent agreement.

Table D4. Rater Correlations and Percentages of Agreement: English II

Score	Teacher vs. ESC				Teacher vs. TEA1				ESC vs. TEA1				TEA1 vs. TEA2			
	N	Cor	EA (%)	EAA (%)	N	Cor	EA (%)	EAA (%)	N	Cor	EA (%)	EAA (%)	N	Cor	EA (%)	EAA (%)
TS1	225	0.40	49	92	224	0.41	44	91	240	0.48	44	93	55	0.70	56	96
PS1	106	0.56	41	90	106	0.49	42	82	242	0.65	48	95	55	0.58	51	93
PS2	36	0.54	42	100	36	0.47	61	94	37	0.49	49	97				
PS3	73	0.60	48	96	73	0.65	41	95	201	0.54	43	96	33	0.92	73	100
TS2	107	0.60	53	96	107	0.57	50	93	242	0.50	44	95	56	0.76	61	98
PF1	237	0.31	35	83	237	0.36	38	89	239	0.43	38	87	66	0.37	41	91
PF2	238	0.23	44	88	237	0.36	45	93	240	0.22	36	86	66	0.31	33	92
PF3	238	0.23	38	84	237	0.25	41	85	240	0.28	36	88	66	0.43	47	94
PF4	236	0.23	33	84	237	0.28	39	85	238	0.32	40	88	66	0.58	47	95
PF5	237	0.49	46	92	237	0.37	44	93	239	0.46	45	94	66	0.50	44	97
PF6	238	0.34	44	87	237	0.47	41	92	240	0.42	44	93	66	0.50	38	95
PF7	237	0.45	41	91	237	0.41	37	89	239	0.53	49	96	66	0.58	39	98
Mean	184	0.41	43	90	184	0.42	44	90	220	0.44	43	92	60	0.57	48	96
SD	79	0.14	6	5	78	0.11	7	4	59	0.12	5	4	10	0.18	11	3
Max	238	0.60	53	100	237	0.65	61	95	242	0.65	49	97	66	0.92	73	100
Min	36	0.23	33	83	36	0.25	37	82	37	0.22	36	86	33	0.31	33	91

Note.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

SD=Standard Deviation, Cor=Correlation, EA=Percentage of exact agreement, EAA=Percentage of exact or adjacent agreement.

Table D5. Summary of Rater Correlations and Percentages of Agreement by Class: Grade 4 Writing

Score	Stat	Teacher vs. ESC				Teacher vs. TEA1				ESC vs. TEA1			
		N	Cor	EA(%)	EAA	N	Cor	EA(%)	EAA	N	Cor	EA(%)	EAA
TS	N	6	6	6	6	7	7	7	7	7	7	7	7
	Mean	64	0.63	48	91	59	0.47	39	88	69	0.51	53	94
	SD	20	0.14	7	5	22	0.22	8	7	22	0.12	7	5
	Max	83	0.80	56	97	84	0.80	55	96	97	0.66	63	99
	Min	33	0.41	39	82	32	0.14	31	75	35	0.31	44	86
PS	N	6	6	6	6	7	7	7	7	7	7	7	7
	Mean	65	0.55	42	89	60	0.54	39	88	70	0.52	47	92
	SD	19	0.16	5	7	21	0.14	12	7	21	0.15	10	5
	Max	83	0.75	47	95	84	0.71	57	96	98	0.77	61	100
	Min	36	0.35	36	75	34	0.35	26	78	36	0.36	36	83
PF1	N	4	4	4	4	4	4	4	4	4	4	4	4
	Mean	43	0.43	38	93	42	0.35	44	93	43	0.05	29	81
	SD	5	0.30	6	4	5	0.23	10	7	5	0.20	13	11
	Max	49	0.66	45	96	49	0.53	55	100	49	0.35	48	98
	Min	39	-0.01	31	88	38	0.02	33	84	38	-0.09	18	73
PF2	N	4	4	4	4	4	4	4	4	4	4	4	4
	Mean	43	0.33	42	91	42	0.36	50	90	43	0.18	37	85
	SD	5	0.13	6	1	5	0.18	12	5	5	0.05	7	5
	Max	49	0.46	48	93	49	0.50	58	95	49	0.25	45	90
	Min	39	0.21	33	90	38	0.09	33	84	38	0.14	29	80
PF3	N	4	4	4	4	4	4	4	4	4	4	4	4
	Mean	43	0.16	39	85	42	0.30	42	92	43	0.22	43	86
	SD	5	0.15	6	4	5	0.20	12	6	5	0.16	5	3
	Max	49	0.35	43	88	49	0.57	58	98	49	0.38	49	90
	Min	39	0.01	30	81	38	0.10	29	84	38	0.03	37	83
PF4	N	4	4	4	4	4	4	4	4	4	4	4	4
	Mean	42	0.24	47	87	42	0.38	43	94	42	0.15	35	84
	SD	4	0.34	17	4	5	0.10	8	7	4	0.13	7	2
	Max	48	0.69	61	93	49	0.50	50	100	48	0.28	44	85
	Min	39	-0.05	23	82	38	0.30	32	84	38	-0.03	28	82
PF5	N	4	4	4	4	4	4	4	4	4	4	4	4
	Mean	42	0.54	42	90	42	0.32	39	90	42	0.46	47	93
	SD	4	0.11	2	7	5	0.15	4	6	5	0.36	16	9
	Max	48	0.68	45	100	49	0.44	43	98	48	0.74	62	100
	Min	38	0.41	39	83	38	0.10	33	84	37	-0.02	24	81
PF6	N	4	4	4	4	4	4	4	4	4	4	4	4
	Mean	42	0.41	40	87	42	0.36	39	90	42	0.38	41	90
	SD	5	0.11	8	7	5	0.20	5	3	5	0.09	8	3
	Max	49	0.55	49	95	49	0.49	43	93	49	0.47	49	93
	Min	38	0.29	32	80	38	0.06	32	86	37	0.28	30	86
PF7	N	4	4	4	4	4	4	4	4	4	4	4	4
	Mean	42	0.66	49	94	42	0.53	46	95	42	0.53	41	94
	SD	5	0.08	3	2	5	0.23	15	3	5	0.12	16	5
	Max	49	0.73	53	95	49	0.73	68	98	49	0.69	56	98
	Min	38	0.57	45	92	38	0.21	34	92	37	0.40	21	86

Note.

TS=Two Time Samples, PS=Three Process Sample, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions. SD=Standard Deviation, Cor=Correlation, EA=Percentage of exact agreement, EAA=Percentage of exact or adjacent agreement

Table D6. Summary of Rater Correlations and Percentages of Agreement by Class: Grade 7 Writing

Score	Stat	Teacher vs. ESC				Teacher vs. TEA1				ESC vs. TEA1			
		N	Cor	EA(%)	EAA	N	Cor	EA(%)	EAA	N	Cor	EA(%)	EAA
TS	N	4	4	4	4	4	4	4	4	4	4	4	4
	Mean	98	0.49	45	92	100	0.60	51	96	98	0.54	50	92
	SD	60	0.04	4	5	59	0.11	10	4	60	0.17	10	6
	Max	185	0.53	50	97	186	0.69	64	100	185	0.74	58	100
	Min	50	0.45	41	86	51	0.45	42	90	50	0.39	37	87
PS	N	4	4	4	4	4	4	4	4	4	4	4	4
	Mean	99	0.52	46	93	101	0.58	47	94	99	0.48	47	90
	SD	60	0.05	4	4	59	0.14	7	4	60	0.12	6	3
	Max	186	0.57	51	96	186	0.70	53	97	186	0.65	55	94
	Min	50	0.47	42	87	52	0.39	37	89	50	0.39	42	87
PF1	N	3	3	3	3	3	3	3	3	3	3	3	3
	Mean	58	0.41	46	88	57	-0.02	55	94	57	0.46	45	88
	SD	30	0.27	5	10	30	0.86	8	5	30	0.05	12	4
	Max	92	0.71	51	96	92	0.68	63	98	92	0.51	53	91
	Min	35	0.18	42	77	35	-0.98	48	89	35	0.41	31	83
PF2	N	3	3	3	3	3	3	3	3	3	3	3	3
	Mean	57	0.34	48	91	57	0.33	47	92	57	0.27	36	90
	SD	30	0.42	2	4	30	0.27	4	7	31	0.13	4	1
	Max	92	0.74	50	96	92	0.54	51	97	92	0.39	39	91
	Min	35	-0.10	46	89	35	0.03	43	84	35	0.13	31	89
PF3	N	3	3	3	3	3	3	3	3	3	3	3	3
	Mean	57	0.46	32	87	57	0.36	45	91	57	0.35	30	83
	SD	31	0.10	11	3	30	0.39	4	3	31	0.07	4	3
	Max	92	0.57	41	91	92	0.75	49	93	92	0.41	34	86
	Min	35	0.38	20	86	35	-0.03	40	87	35	0.28	26	80
PF4	N	3	3	3	3	3	3	3	3	3	3	3	3
	Mean	57	0.40	33	85	57	0.40	37	89	57	0.28	33	80
	SD	30	0.19	8	4	30	0.36	13	8	30	0.18	7	3
	Max	91	0.59	41	87	92	0.74	46	94	91	0.49	40	82
	Min	35	0.22	26	80	35	0.02	22	80	35	0.15	26	77
PF5	N	3	3	3	3	3	3	3	3	3	3	3	3
	Mean	58	0.51	42	87	57	0.58	56	98	57	0.45	42	87
	SD	30	0.21	7	9	30	0.19	4	2	30	0.04	12	5
	Max	92	0.71	46	96	92	0.76	60	100	92	0.50	54	92
	Min	35	0.30	34	77	35	0.39	52	96	35	0.41	31	83
PF6	N	3	3	3	3	3	3	3	3	3	3	3	3
	Mean	58	0.40	44	88	57	0.47	49	98	57	0.38	39	86
	SD	30	0.37	6	7	30	0.19	2	0	30	0.08	9	8
	Max	92	0.75	50	93	92	0.61	51	98	92	0.46	49	91
	Min	35	0.01	37	80	35	0.26	48	97	35	0.30	31	77
PF7	N	3	3	3	3	3	3	3	3	3	3	3	3
	Mean	57	0.59	47	92	57	0.54	57	96	57	0.45	34	89
	SD	30	0.20	4	5	30	0.05	12	1	30	0.15	6	4
	Max	91	0.76	50	98	91	0.57	66	97	92	0.61	40	93
	Min	35	0.37	42	89	35	0.48	43	96	35	0.31	29	86

Note.

TS=Two Time Samples, PS=Three Process Sample, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions. SD=Standard Deviation, Cor=Correlation, EA=Percentage of exact agreement, EAA=Percentage of exact or adjacent agreement

Table D7. Summary of Rater Correlations and Percentages of Agreement by Class: English I

Score	Stat	Teacher vs. ESC				Teacher vs. TEA1				ESC vs. TEA1			
		N	Cor	EA(%)	EAA	N	Cor	EA(%)	EAA	N	Cor	EA(%)	EAA
TS	N	4	4	4	4	4	4	4	4	4	4	4	4
	Mean	51	0.59	50	94	51	0.54	48	93	72	0.48	47	92
	SD	22	0.29	8	8	22	0.22	12	6	40	0.06	4	4
	Max	84	0.88	58	100	84	0.70	61	98	123	0.57	51	95
	Min	36	0.20	39	82	36	0.22	33	83	40	0.43	43	88
PS	N	3	3	3	3	3	3	3	3	4	4	4	4
	Mean	55	0.56	46	91	55	0.50	46	89	72	0.45	46	91
	SD	25	0.08	12	2	25	0.07	9	3	40	0.06	6	3
	Max	84	0.63	53	93	84	0.54	53	93	122	0.53	53	94
	Min	40	0.48	33	90	40	0.42	35	88	40	0.39	40	88
PF1	N	2	2	2	2	2	2	2	2	2	2	2	2
	Mean	52	0.38	22	72	52	0.39	28	67	52	0.32	43	88
	SD	14	0.24	14	21	13	0.13	23	35	13	0.00	6	3
	Max	62	0.55	32	87	61	0.48	44	92	61	0.32	48	90
	Min	42	0.21	12	57	42	0.31	12	43	42	0.31	39	86
PF2	N	2	2	2	2	2	2	2	2	2	2	2	2
	Mean	52	0.44	34	87	52	0.61	42	90	51	0.17	38	87
	SD	13	0.03	8	9	13	0.04	19	9	13	0.03	6	1
	Max	61	0.46	39	93	61	0.64	56	97	60	0.19	42	88
	Min	42	0.42	29	81	42	0.59	29	83	42	0.15	33	87
PF3	N	2	2	2	2	2	2	2	2	2	2	2	2
	Mean	52	0.21	33	83	52	0.36	37	91	52	0.19	34	85
	SD	14	0.07	1	1	13	0.04	1	4	13	0.11	5	1
	Max	62	0.26	33	83	61	0.39	38	93	61	0.26	38	86
	Min	42	0.17	32	82	42	0.34	36	88	42	0.11	31	84
PF4	N	2	2	2	2	2	2	2	2	2	2	2	2
	Mean	52	0.19	43	87	52	0.45	40	94	52	0.19	35	84
	SD	14	0.03	3	2	13	0.02	6	4	13	0.16	1	1
	Max	62	0.21	45	88	61	0.47	44	97	61	0.31	36	85
	Min	42	0.16	40	85	42	0.44	36	90	42	0.08	34	83
PF5	N	2	2	2	2	2	2	2	2	2	2	2	2
	Mean	52	0.55	49	96	52	0.47	34	89	51	0.30	36	93
	SD	13	0.03	2	1	13	0.01	5	11	13	0.30	17	10
	Max	61	0.57	50	97	61	0.48	38	97	60	0.51	48	100
	Min	42	0.53	48	95	42	0.46	31	81	42	0.08	24	86
PF6	N	2	2	2	2	2	2	2	2	2	2	2	2
	Mean	52	0.38	44	93	52	0.55	46	90	52	0.19	38	90
	SD	14	0.05	2	3	13	0.04	11	9	13	0.45	4	12
	Max	62	0.42	45	95	61	0.58	54	97	61	0.51	41	98
	Min	42	0.35	43	90	42	0.52	38	83	42	-0.13	36	81
PF7	N	2	2	2	2	2	2	2	2	2	2	2	2
	Mean	52	0.35	46	93	52	0.53	44	94	52	0.37	48	94
	SD	14	0.20	10	7	13	0.14	8	4	13	0.22	7	2
	Max	62	0.49	52	98	61	0.63	49	97	61	0.53	52	95
	Min	42	0.20	39	88	42	0.42	38	90	42	0.21	43	93

Note.

TS=Two Time Samples, PS=Three Process Sample, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions. SD=Standard Deviation, Cor=Correlation, EA=Percentage of exact agreement, EAA=Percentage of exact or adjacent agreement.

Table D8. Summary of Rater Correlations and Percentages of Agreement by Class: English II

Score	Stat	Teacher vs. ESC				Teacher vs. TEA1				ESC vs. TEA1			
		N	Cor	EA(%)	EAA	N	Cor	EA(%)	EAA	N	Cor	EA(%)	EAA
TS	N	6	6	6	6	6	6	6	6	6	6	6	6
	Mean	55	0.42	52	94	55	0.43	46	91	80	0.43	45	94
	SD	16	0.16	9	3	16	0.19	14	6	45	0.10	7	3
	Max	78	0.62	67	98	78	0.75	67	98	160	0.59	55	98
	Min	37	0.23	39	91	37	0.22	23	81	42	0.27	38	89
PS	N	4	4	4	4	4	4	4	4	6	6	6	6
	Mean	53	0.34	43	93	53	0.39	44	87	79	0.53	46	95
	SD	12	0.40	9	6	12	0.43	15	11	45	0.07	4	2
	Max	69	0.59	52	99	69	0.78	57	96	160	0.59	52	97
	Min	42	-0.26	31	88	42	-0.21	25	71	42	0.44	40	93
PF1	N	3	3	3	3	3	3	3	3	3	3	3	3
	Mean	54	0.32	35	85	55	0.45	41	93	55	0.49	44	90
	SD	22	0.11	7	10	22	0.14	8	6	22	0.09	13	2
	Max	79	0.43	39	94	79	0.58	49	100	79	0.59	59	92
	Min	36	0.21	27	75	36	0.31	34	87	37	0.43	36	89
PF2	N	3	3	3	3	3	3	3	3	3	3	3	3
	Mean	55	0.19	47	86	55	0.58	51	95	56	0.21	36	86
	SD	22	0.23	11	10	22	0.26	14	5	21	0.08	5	2
	Max	79	0.44	56	95	79	0.82	64	100	79	0.26	41	89
	Min	36	-0.02	35	76	36	0.31	35	90	37	0.12	31	84
PF3	N	3	3	3	3	3	3	3	3	3	3	3	3
	Mean	55	0.11	40	80	55	0.31	45	87	56	0.24	36	90
	SD	22	0.26	20	12	22	0.14	4	10	21	0.16	1	1
	Max	79	0.32	55	90	79	0.43	49	94	79	0.37	37	91
	Min	36	-0.19	17	67	36	0.16	42	75	37	0.07	35	88
PF4	N	3	3	3	3	3	3	3	3	3	3	3	3
	Mean	54	0.17	33	81	55	0.34	44	89	55	0.37	40	89
	SD	22	0.29	14	7	22	0.07	5	5	22	0.15	6	7
	Max	79	0.50	41	87	79	0.39	47	94	79	0.54	45	94
	Min	35	-0.01	17	74	36	0.26	39	83	36	0.27	33	81
PF5	N	3	3	3	3	3	3	3	3	3	3	3	3
	Mean	55	0.51	45	92	55	0.42	45	92	56	0.35	44	92
	SD	22	0.13	7	5	22	0.04	11	1	21	0.14	1	5
	Max	79	0.61	53	98	79	0.44	53	94	79	0.50	46	97
	Min	36	0.36	39	89	36	0.38	33	91	37	0.23	43	89
PF6	N	3	3	3	3	3	3	3	3	3	3	3	3
	Mean	55	0.30	39	81	55	0.51	38	89	56	0.26	42	91
	SD	22	0.16	12	13	22	0.05	12	5	21	0.10	1	4
	Max	79	0.44	53	90	79	0.56	51	92	79	0.35	43	94
	Min	36	0.12	31	67	36	0.45	28	83	37	0.14	41	86
PF7	N	3	3	3	3	3	3	3	3	3	3	3	3
	Mean	54	0.48	37	90	55	0.50	36	88	55	0.46	48	98
	SD	22	0.06	13	8	22	0.11	11	9	22	0.07	6	1
	Max	79	0.54	52	94	79	0.61	45	94	79	0.54	52	100
	Min	36	0.41	28	81	36	0.39	24	78	37	0.41	42	97

Note.

TS=Two Time Samples, PS=Three Process Sample, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions. SD=Standard Deviation, Cor=Correlation, EA=Percentage of exact agreement, EAA=Percentage of exact or adjacent agreement

Appendix E: Rater Score Reliability

The rater reliability was calculated based on the generalizability theory (Brennan, 2001[3]; Shavelson & Webb, 1991[4]). In particular, for each rating score we fitted the model P × R for the G study and the model P × r for the D study, and for the sum of the seven portfolio scores we fitted the model P × R × S for the G study and the model P × r × S for the D study, where P refers to students, R denotes three raters (teacher, ESC, TEA1), S denotes the seven portfolio scores, and r denotes one or two raters in the D study. S was treated as a fixed effect while the others were treated as random effects. Generalizability coefficients are calculated for one and two raters as the reliability indicator. The generalizability coefficient is analogous to the Cronbach's alpha in the classical theory.

Rater score reliabilities were calculated for each rating score and the sum of the seven portfolio rating scores for one rater and two raters at the class level. We used teacher, ESC, and TEA1 rating scores for the reliability calculations, and only the classes with sample sizes of at least 30 were included. Tables E1–E4 summarize the reliabilities across classes in the four tests, respectively. Figure E1 below presents the mean reliabilities across classes in the four tests. Across the four tests and all rating scores, the mean reliabilities over the classes ranged from 0.17 to 0.49 for one rater and from 0.27 to 0.66 for sum of two raters; the maximum reliability was 0.65 for one rater and 0.79 for sum of two raters.

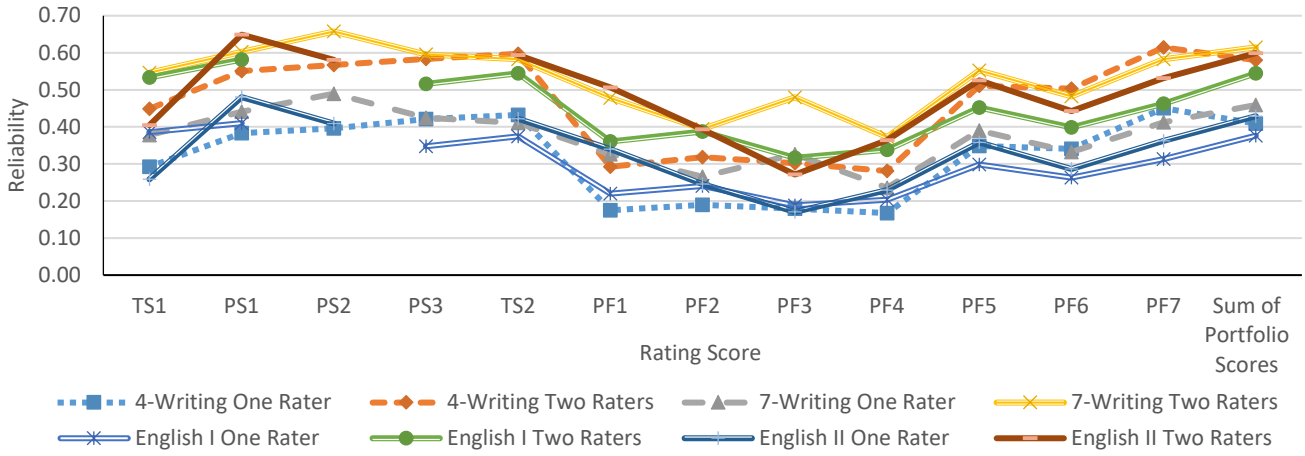


Figure E1. Mean reliability of rating scores over classes for one and two raters.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions

[3] Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

[4] Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Generally speaking, test scores should have a reliability of at least .90 if used for high-stakes purposes and at least 0.80 or 0.85 for low-stake purposes (Wells & Wollack, 2003). For a reliability of 0.80 the inter-rater correlation needs to be 0.89. No score in the current study met this reliability criterion even if two raters and the sum of their scores were used as the reporting score. Most of the scores were far below this criterion. As a reference for interpretation, the reliabilities for spring 2017 STAAR grades 4 and 7 summative writing tests were 0.84 and 0.86 respectively. Since English I and English II assess both reading and writing, they are not provided as a comparison framework.

Table E1. Rater Score Reliability: Grade 4 Writing

Score	N Classes	N				One Rater				Two Raters			
		Mean	SD	Max	Min	Mean	SD	Max	Min	Mean	SD	Max	Min
TS1	4	42	4	48	39	0.29	0.08	0.41	0.23	0.45	0.09	0.58	0.37
PS1	5	40	7	49	30	0.38	0.08	0.52	0.30	0.55	0.08	0.68	0.46
PS2	1	40		40	40	0.40		0.40	0.40	0.57		0.57	0.57
PS3	3	37	6	42	30	0.42	0.14	0.53	0.26	0.58	0.15	0.69	0.42
TS2	4	37	5	42	30	0.43	0.11	0.60	0.35	0.60	0.10	0.75	0.52
PF1	5	40	7	49	30	0.17	0.07	0.25	0.10	0.29	0.10	0.40	0.17
PF2	5	40	7	49	30	0.19	0.04	0.22	0.13	0.32	0.05	0.36	0.23
PF3	4	42	5	49	38	0.18	0.04	0.24	0.15	0.30	0.06	0.39	0.26
PF4	4	42	5	48	38	0.17	0.08	0.27	0.10	0.28	0.11	0.43	0.18
PF5	5	39	7	48	30	0.35	0.10	0.48	0.21	0.51	0.11	0.65	0.35
PF6	5	40	7	49	30	0.34	0.09	0.47	0.22	0.50	0.10	0.64	0.37
PF7	5	40	7	49	30	0.45	0.11	0.58	0.31	0.61	0.11	0.73	0.48
Sum of Portfolio Scores	4	41	5	47	36	0.41	0.03	0.44	0.38	0.58	0.03	0.61	0.55

Note.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

SD=Standard Deviation.

Table E2. Rater Score Reliability: Grade 7 Writing

Score	N Classes	N				One Rater				Two Raters			
		Mean	SD	Max	Min	Mean	SD	Max	Min	Mean	SD	Max	Min
TS1	3	57	31	92	37	0.38	0.06	0.44	0.33	0.55	0.06	0.61	0.50
PS1	3	57	31	93	36	0.44	0.15	0.58	0.29	0.60	0.14	0.74	0.45
PS2	1	42		42	42	0.49		0.49	0.49	0.66		0.66	0.66
PS3	2	65	40	93	37	0.42	0.05	0.46	0.39	0.60	0.05	0.63	0.56
TS2	3	57	31	93	37	0.41	0.03	0.43	0.38	0.58	0.03	0.60	0.55
PF1	3	57	30	92	35	0.33	0.16	0.51	0.21	0.48	0.18	0.68	0.34
PF2	3	57	31	92	35	0.27	0.19	0.46	0.08	0.39	0.24	0.63	0.15
PF3	3	57	31	92	35	0.33	0.15	0.49	0.21	0.48	0.16	0.66	0.35
PF4	3	57	30	91	35	0.24	0.11	0.32	0.11	0.37	0.15	0.48	0.20
PF5	3	57	30	92	35	0.39	0.13	0.51	0.25	0.55	0.14	0.67	0.40
PF6	3	57	30	92	35	0.33	0.16	0.49	0.16	0.48	0.19	0.65	0.28
PF7	3	57	30	91	35	0.41	0.05	0.46	0.36	0.58	0.05	0.63	0.53
Sum of Portfolio Scores	3	56	30	90	35	0.46	0.18	0.65	0.29	0.61	0.17	0.79	0.45

Note.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

SD=Standard Deviation.

Table E3. Rater Score Reliability: English I

Score	N Classes	N				One Rater				Two Raters			
		Mean	SD	Max	Min	Mean	SD	Max	Min	Mean	SD	Max	Min
TS1	2	42	0	42	42	0.39	0.24	0.56	0.21	0.53	0.26	0.72	0.35
PS1	1	42		42	42	0.41		0.41	0.41	0.58		0.58	0.58
PS3	1	42		42	42	0.35		0.35	0.35	0.52		0.52	0.52
TS2	1	42		42	42	0.37		0.37	0.37	0.54		0.54	0.54
PF1	2	52	13	61	42	0.22	0.01	0.22	0.22	0.36	0.01	0.37	0.35
PF2	2	51	13	60	42	0.24	0.03	0.27	0.22	0.39	0.04	0.42	0.36
PF3	2	52	13	61	42	0.19	0.02	0.20	0.17	0.32	0.03	0.34	0.30
PF4	2	52	13	61	42	0.20	0.02	0.22	0.19	0.34	0.03	0.36	0.32
PF5	2	51	13	60	42	0.30	0.13	0.39	0.21	0.45	0.15	0.56	0.35
PF6	2	52	13	61	42	0.26	0.19	0.40	0.13	0.40	0.24	0.57	0.23
PF7	2	52	13	61	42	0.31	0.18	0.44	0.19	0.46	0.21	0.61	0.31
Sum of Portfolio Scores	2	51	13	60	42	0.38	0.07	0.43	0.32	0.54	0.08	0.60	0.49

Note.

TS1=Time Sample 1, PS1=Process Sample 1, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

SD=Standard Deviation.

Table E4. Rater Score Reliability: English II

Score	N Classes	N				One Rater				Two Raters			
		Mean	SD	Max	Min	Mean	SD	Max	Min	Mean	SD	Max	Min
TS1	3	51	24	78	37	0.26	0.10	0.36	0.16	0.40	0.13	0.52	0.27
PS1	1	33		33	33	0.48		0.48	0.48	0.65		0.65	0.65
PS2	1	36		36	36	0.41		0.41	0.41	0.58		0.58	0.58
TS2	1	34		34	34	0.42		0.42	0.42	0.59		0.59	0.59
PF1	3	54	22	79	36	0.34	0.03	0.38	0.32	0.51	0.03	0.55	0.49
PF2	3	55	22	79	36	0.25	0.04	0.27	0.20	0.39	0.05	0.42	0.33
PF3	3	55	22	79	36	0.17	0.15	0.27	0.00	0.27	0.24	0.42	0.00
PF4	3	54	22	79	35	0.23	0.11	0.34	0.12	0.36	0.15	0.51	0.21
PF5	3	55	22	79	36	0.36	0.05	0.40	0.31	0.53	0.05	0.58	0.47
PF6	3	55	22	79	36	0.29	0.06	0.36	0.24	0.44	0.07	0.53	0.39
PF7	3	54	22	79	36	0.36	0.02	0.38	0.34	0.53	0.03	0.55	0.50
Sum of Portfolio Scores	3	54	23	79	35	0.43	0.06	0.49	0.37	0.60	0.06	0.66	0.54

Note.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

SD=Standard Deviation.

Appendix F: Correlations of Portfolio Scores

To check whether the separation of the seven portfolio scores is justifiable, the polychoric correlations were calculated among the seven portfolio scores for each rater group. The polychoric correlation was introduced previously in Appendix D.

Tables F1–F4 show the correlations for the four tests, respectively. In general, one can make the following observations:

1. Within each of the following groups, scores were highly correlated (ranged from 0.72 to 0.97) and distinct from the other portfolio scores across all subjects and rater groups:
    a. the last three portfolio scores (i.e., Expressing Ideas, Organization and Structure, and Use of Language and Conventions),
    b. the first and second portfolio scores (i.e., Planning and Drafting),
    c. the third and fourth portfolio scores (i.e., Revising and Editing, Publishing and Attention to Feedback).
2. Except the three score groups mentioned in Point 1, the correlations of the portfolio scores by teachers were, in general, higher than those by ESC and TEA1 raters across tests, except for TEA1 raters on grade 4 where the correlations by TEA1 raters were similar to those by teachers. Most of those correlations by teachers were higher than 0.70, while most of the correlations by ESC and TEA1 raters were below 0.70.

Because most of the correlations were intermediate to high (i.e., between 0.45 and 0.90), the separation of the seven portfolio scores were in general justified by the data. They were not too high (i.e., higher than 0.90) to indicate that any two areas of the portfolio rubric cannot be distinguished, nor were they too low to indicate that any two areas of the portfolio rubric are basically not related to each other that is contradictive to the underlying writing theory.

Table F1. Correlations of Portfolio Scores: Grade 4 Writing

Rater	Score	PF1	PF2	PF3	PF4	PF5	PF6
Teacher	PF2	0.87					
	PF3	0.77	0.77				
	PF4	0.76	0.74	0.86			
	PF5	0.71	0.66	0.77	0.75		
	PF6	0.67	0.64	0.75	0.74	0.90	
	PF7	0.70	0.66	0.71	0.80	0.88	0.88
ESC	PF2	0.78					
	PF3	0.49	0.59				
	PF4	0.47	0.55	0.81			
	PF5	0.50	0.60	0.49	0.56		
	PF6	0.52	0.59	0.49	0.53	0.90	
	PF7	0.49	0.56	0.50	0.46	0.84	0.85
TEA1	PF2	0.81					
	PF3	0.63	0.70				
	PF4	0.66	0.70	0.91			
	PF5	0.68	0.68	0.67	0.74		
	PF6	0.69	0.74	0.72	0.75	0.85	
	PF7	0.66	0.70	0.67	0.70	0.85	0.88

Note. PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

Table F2. Correlations of Portfolio Scores: Grade 7 Writing

Rater	Score	PF1	PF2	PF3	PF4	PF5	PF6
Teacher	PF2	0.87					
	PF3	0.92	0.94				
	PF4	0.85	0.90	0.95			
	PF5	0.81	0.82	0.85	0.84		
	PF6	0.76	0.90	0.87	0.88	0.94	
	PF7	0.78	0.88	0.86	0.94	0.93	0.95
ESC	PF2	0.88					
	PF3	0.62	0.72				
	PF4	0.54	0.59	0.85			
	PF5	0.60	0.62	0.57	0.59		
	PF6	0.61	0.67	0.60	0.67	0.87	
	PF7	0.62	0.68	0.58	0.64	0.87	0.83
TEA1	PF2	0.89					
	PF3	0.71	0.78				
	PF4	0.64	0.77	0.93			
	PF5	0.47	0.59	0.59	0.65		
	PF6	0.58	0.59	0.56	0.62	0.94	
	PF7	0.50	0.62	0.60	0.69	0.91	0.89

Note. PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

Table F3. Correlations of Portfolio Scores: English I

Rater	Score	PF1	PF2	PF3	PF4	PF5	PF6
Teacher	PF2	0.83					
	PF3	0.70	0.78				
	PF4	0.45	0.67	0.89			
	PF5	0.64	0.83	0.82	0.75		
	PF6	0.53	0.78	0.81	0.74	0.96	
	PF7	0.56	0.76	0.78	0.72	0.93	0.97
ESC	PF2	0.78					
	PF3	0.50	0.69				
	PF4	0.50	0.65	0.90			
	PF5	0.48	0.55	0.58	0.65		
	PF6	0.46	0.57	0.63	0.62	0.84	
	PF7	0.47	0.53	0.62	0.65	0.79	0.79
TEA1	PF2	0.75					
	PF3	0.62	0.73				
	PF4	0.56	0.64	0.86			
	PF5	0.65	0.65	0.54	0.59		
	PF6	0.62	0.74	0.62	0.61	0.93	
	PF7	0.60	0.62	0.56	0.61	0.81	0.87

Note. PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

Table F4. Correlations of Portfolio Scores: English II

Rater	Score	PF1	PF2	PF3	PF4	PF5	PF6
Teacher	PF2	0.85					
	PF3	0.72	0.77				
	PF4	0.78	0.79	0.91			
	PF5	0.70	0.81	0.66	0.75		
	PF6	0.73	0.76	0.71	0.75	0.86	
	PF7	0.74	0.85	0.68	0.79	0.89	0.89
ESC	PF2	0.72					
	PF3	0.52	0.60				
	PF4	0.46	0.64	0.80			
	PF5	0.48	0.57	0.50	0.57		
	PF6	0.42	0.59	0.49	0.55	0.80	
	PF7	0.45	0.54	0.53	0.64	0.86	0.81
TEA1	PF2	0.77					
	PF3	0.58	0.66				
	PF4	0.57	0.68	0.88			
	PF5	0.49	0.61	0.64	0.70		
	PF6	0.56	0.67	0.64	0.73	0.92	
	PF7	0.56	0.67	0.61	0.76	0.83	0.91

Note. PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

Appendix G: Correlations between Writing Pilot Scores and STAAR Summative Writing Scores

The polyserial correlations were calculated between the 12 rating scores and the corresponding spring 2017 STAAR writing scale scores for each test and rater group. Also computed is the Pearson correlation between the sum of the portfolio scores and the spring 2017 STAAR writing scale score for each test and rater group. A sample size of at least 30 was required for each calculation. This correlation can serve as an external validity indicator for a rater score.

Polyserial correlation (Drasgow, 1988) is appropriate for the case where one variable is an ordered categorical variable and the other is a continuous variable. Like polychoric correlation, polyserial correlation assumes a continuous variable underlying the categorical variable and the two continuous variables follow a binormal distribution. Polyserial correlation is estimated by the maximum likelihood estimation. If the assumptions hold, polyserial correlation more accurately reflects the association between one ordered categorical variable and one continuous variable, while Pearson correlation tends to underestimate the association. For the sums of the seven portfolio scores, their correlations are Pearson correlations because both variables are considered to be continuous.

Tables G1–G4 list the correlations for all rating scores and rater groups in the four tests, respectively. The correlations are also plotted in Figures G1–G4 for easy observation. In most cases, the correlations were low to intermediate (from 0.30 to 0.65). The correlations for teachers in grade 7 writing were relatively higher in the range from 0.50 to 0.85, while some portfolio scores by teachers in grade 4 writing, English I, and English II and by ESC raters in English II had the correlations below 0.30. Therefore, most of the rating scores correlated with the STAAR writing scale scores to some extent, while for some rating scores the correlations were very weak.

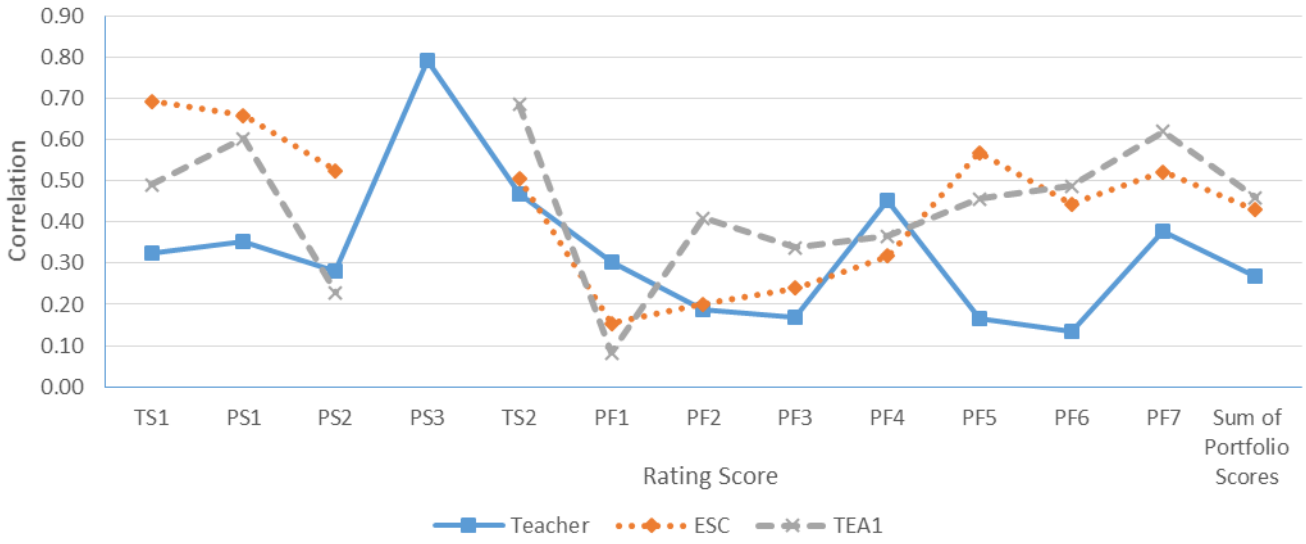


Figure G1. Correlation between grade 4 writing rating scores and spring 2017 STAAR scale scores.
TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

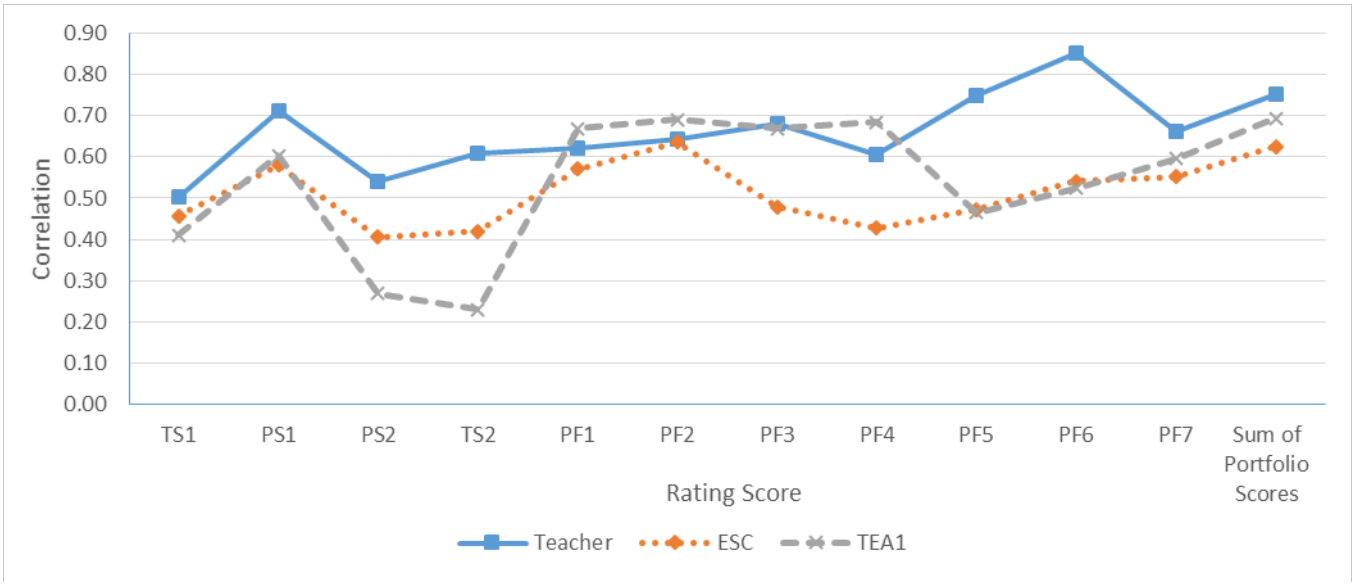


Figure G2. Correlation between grade 7 writing rating scores and spring 2017 STAAR scale scores.
TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

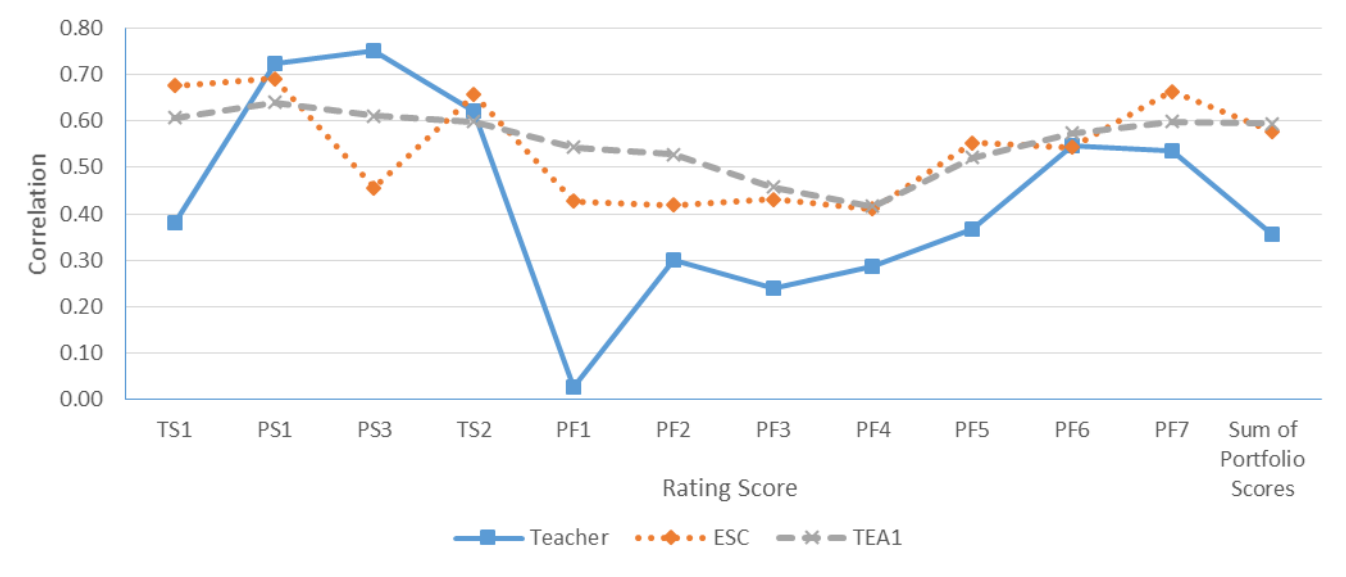


Figure G3. Correlation between English I rating scores and spring 2017 STAAR scale scores.

TS1=Time Sample 1, PS1=Process Sample 1, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

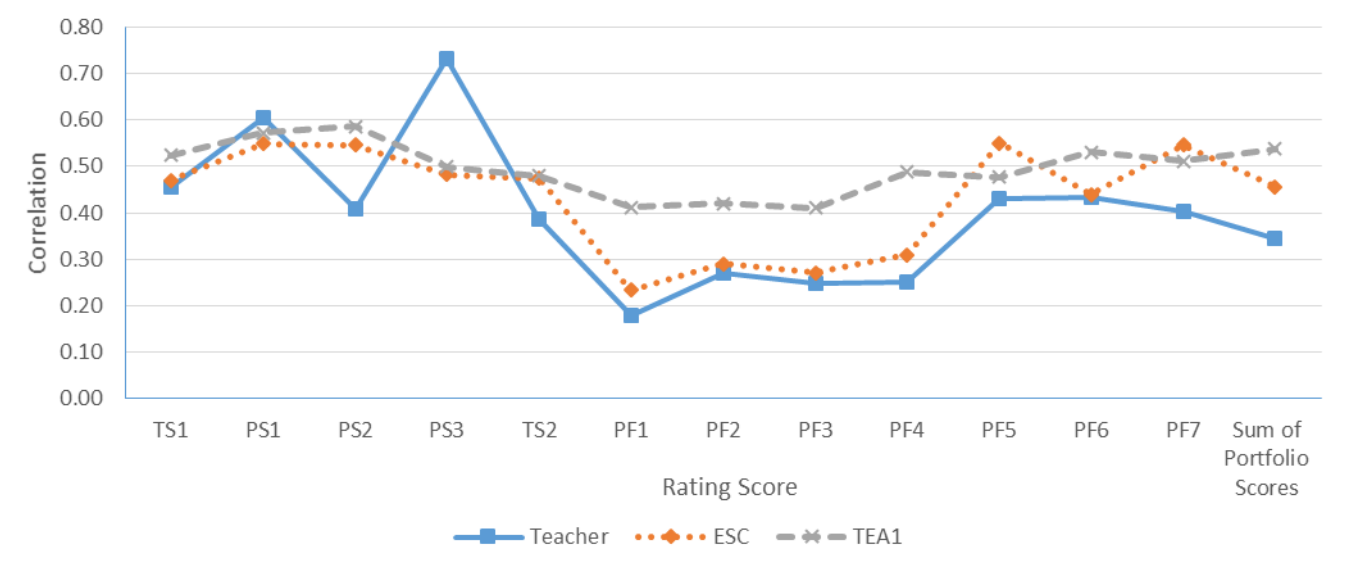


Figure G4. Correlation between English II rating scores and spring 2017 STAAR scale scores.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

Table G1. Correlations between Rating Scores and Spring 2017 STAAR Scale Scores: Grade 4 Writing

Score	Teacher		ESC		TEA1	
	N	Cor	N	Cor	N	Cor
TS1	111	0.32	65	0.69	64	0.49
PS1	106	0.35	65	0.66	65	0.60
PS2	104	0.28	64	0.52	64	0.23
PS3	59	0.79				
TS2	95	0.47	61	0.50	61	0.68
PF1	64	0.30	66	0.15	66	0.08
PF2	64	0.19	65	0.20	66	0.41
PF3	64	0.17	64	0.24	66	0.34
PF4	64	0.45	63	0.32	66	0.37
PF5	64	0.17	66	0.57	66	0.46
PF6	64	0.13	66	0.44	66	0.49
PF7	64	0.38	66	0.52	66	0.62
Sum of Portfolio Scores	64	0.27	63	0.43	66	0.46

Note.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

Cor=Correlation.

Table G2. Correlations between Rating Scores and Spring 2017 STAAR Scale Scores: Grade 7 Writing

Score	Teacher		ESC		TEA1	
	N	Cor	N	Cor	N	Cor
TS1	144	0.50	53	0.46	57	0.41
PS1	141	0.71	56	0.58	58	0.60
PS2	136	0.54	55	0.41	58	0.27
TS2	133	0.61	55	0.42	57	0.23
PF1	59	0.62	59	0.57	58	0.67
PF2	59	0.64	58	0.64	58	0.69
PF3	59	0.68	57	0.48	58	0.67
PF4	59	0.60	59	0.43	57	0.68
PF5	59	0.75	59	0.47	58	0.46
PF6	59	0.85	59	0.54	58	0.52
PF7	59	0.66	59	0.55	58	0.60
Sum of Portfolio Scores	59	0.75	57	0.62	57	0.69

Note.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

Cor=Correlation.

Table G3. Correlations between Rating Scores and Spring 2017 STAAR Scale Scores: English I

Score	Teacher		ESC		TEA1	
	N	Cor	N	Cor	N	Cor
TS1	292	0.38	147	0.68	145	0.61
PS1	135	0.72	147	0.69	146	0.64
PS3	95	0.75	128	0.46	125	0.61
TS2	114	0.62	147	0.66	146	0.60
PF1	147	0.03	147	0.43	146	0.54
PF2	147	0.30	146	0.42	146	0.53
PF3	147	0.24	145	0.43	146	0.46
PF4	147	0.29	146	0.41	146	0.42
PF5	147	0.37	146	0.55	146	0.52
PF6	147	0.55	147	0.54	146	0.57
PF7	147	0.53	147	0.66	146	0.60
Sum of Portfolio Scores	147	0.36	144	0.58	146	0.59

Note.

TS1=Time Sample 1, PS1=Process Sample 1, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

Cor=Correlation.

Table G4. Correlations between Rating Scores and Spring 2017 STAAR Scale Scores: English II

Score	Teacher		ESC		TEA1	
	N	Cor	N	Cor	N	Cor
TS1	322	0.45	224	0.47	223	0.52
PS1	122	0.60	225	0.55	225	0.57
PS2	39	0.41	33	0.55	33	0.59
PS3	85	0.73	189	0.48	188	0.50
TS2	119	0.39	225	0.47	225	0.48
PF1	221	0.18	223	0.23	224	0.41
PF2	221	0.27	224	0.29	224	0.42
PF3	221	0.25	224	0.27	224	0.41
PF4	221	0.25	222	0.31	224	0.49
PF5	221	0.43	223	0.55	224	0.48
PF6	221	0.43	224	0.44	224	0.53
PF7	221	0.40	223	0.55	224	0.51
Sum of Portfolio Scores	221	0.34	219	0.46	224	0.54

Note.

TS1=Time Sample 1, PS1=Process Sample 1, PS2=Process Sample 2, PS3=Process Sample 3, TS2=Time Sample 2, PF1=Planning, PF2=Drafting, PF3=Revising, PF4=Editing, Publishing and Attention to Feedback, PF5=Expressing Ideas, PF6=Organization and Structure, PF7=Use of Language and Conventions.

Cor=Correlation.